HUMAN PHYSIOLOGY

BERNARDO A. HOUSSAY, M.D. Professor of Physiology; Director of the Institute of Biology and Experimental Medicine, Buenos Aires, Argentnia

JUAN T. LEWIS, M.D. Professor of Physiology; Director of the Institute for Medical Research, Rosario, Argentina

OSCAR ORÍAS, M.D. Professor of Physiology; Director of the Mercedes and Martín Ferreyra Institute for Medical Research, Córdoba, Argentina

EDUARDO BRAUN-MENÉNDEZ, M.D. Professor of Physiology; Member of the Institute ot Biology and Experimental Medicine, Buenos Aires, Argentina

ENRIQUE HUG, M.D. Professor of Pharmacology of the School of Medicine, Rosario, Argentina

VIRGILIO G. FOGLIA, M.D. Professor of Physiology; Member of the Institute of Biology and Experimental Medicine, Buenos Aires, Argentina

LUIS F. LELOIR, M.D. Director of the Institute for Biochemical Research, Campomar Foundation, Buenos Aires, Argentina

Translated by *JUAN T. LEWIS, M.D.*, and *OLIVE T. LEWIS*

Foreword by HERBERT M. EVANS, M.D.

SECOND EDITION

INTERNATIONAL STUDENT EDITION

McGRAW-HILL BOOK COMPANY, INC.

New York *Toronto* *London*

KŌGAKUSHA COMPANY, LTD.

Tokyo

To the memory of

JUAN BAUTISTA SAUBERAN

and to

THE JUAN BAUTISTA SAUBERAN FOUNDATION

in grateful acknowledgment

of their work in support of

scientific research

this book is dedicated

Foreword

While there is no avowed tendency to "nationalism" in science, it is inevitable that any scientist will give more attention and, hence, more apparent importance to what is being done in his own country. Such a tendency has always been known to be responsible for overemphasis on the laboratory from which emanates an author's or editor's own work. This is, perhaps, an unavoidable tendency, inherent in the nature of things, but it is pernicious because it detracts from an essential attribute of science, its universality. The geographic and, to some extent, also the cultural detachment of the great continent to the south of us has yielded an unique fruit in this remarkable treatise on physiology. It is the product of a group of men who not only are fruitful investigators but who, working where science has not hitherto been intensively cultivated, have felt themselves forced to study conscientiously the literature in mammalian physiology produced throughout the world. They have thus the perspective requisite for a balanced account.

The group has furthermore led an ideal intellectual life with daily or at least frequent intercommunication. It has been led by a great individual, Bernardo A. Houssay, whose capacity to break new ground has been adequately acknowledged by a Nobel award. Those who had watched the initiation of Professor Houssay's career and the remarkable School which he had set up in that great capital, Buenos Aires, were not surprised by his brilliant Dunham Lectures at Harvard University some years ago, but it is perhaps safe to say that the Dunham Lectures first introduced this scholar and his pupils to his North American confreres.

The group effort just mentioned has conferred upon this treatise a remarkable unity, a unity otherwise given a treatise only by a single individual, and it is well known that it is no longer possible for an individual to cover the whole field of a given science. Furthermore, the English version of the treatise has been provided by one of its authors, Juan Lewis, and that has itself aided the uniformity and harmony of the book.

While the sections on circulatory physiology, under the able auspices of Orías, and renal physiology, under Braun-Menéndez, could be compared advantageously with any textual treatments of these themes, it is inevitable that the outstanding excellence of this treatise will be considered its treatment of the field of endocrine physiology, a field of enormous present-day importance, in which almost daily advances are being made. All North American colleagues of these Argentine scientists now welcome warmly this signal accomplishment of the group.

Herbert M. Evans
University of California

Preface

This book has been written for medical students and for doctors of medicine who wish to study the fundamental principles of modern physiology. It has been written for students who have no knowledge of pathology, but who must acquire the physiological basis necessary for the study of disease and the practice of medicine. A book to be used in medical education cannot be merely a textbook of applied medical physiology; therefore the facts of human physiology have been given preferential, but not exclusive, consideration. Human physiology is better understood in the light of general and comparative physiology, and a living organism is more completely known when there is some knowledge of other species. Many advances in human physiology have originated in studies on isolated cells, on yeasts, and on the lower animals.

Physiology is a science that studies the phenomena occurring in living organisms and endeavors to establish their laws. It is that part of biological science which studies the functions of living organisms in health and disease. The knowledge of physiology is indispensable in medicine because the same general laws apply to normal and to pathologic conditions of life; disease is simply the modification or deviation of normal function and cannot be understood without previous knowledge of the normal state.

Physiology as a science is not dependent on medicine, and as such it should be studied for its own sake, without being limited to its medical, veterinary, or zootechnical applications. Even in a medical school, the study of physiology should not be reduced to problems of human physiology and its medical applications. Undoubtedly, however, human medicine has been the most powerful stimulus for the study of physiology, and even now in several countries it is taught only in medical schools.

An outstanding feature of modern medicine is its physiological foundation. Medicine has a threefold purpose: to preserve health, to prevent disease, and to cure disease. The doctor in the first place must know how to preserve and improve health, which is the normal condition of man's life, whereas illness attacks him only during short periods. The advancement of physiology and experimental medicine has been instrumental in the lengthening of life and in ever greater efficiency in the prevention and treatment of disease.

Physiological discoveries sooner or later always find useful applications in medicine and in hygiene, but they must first become familiar to the medical man. Unfortunately, many years often pass by before these applications enter into current medical practice. A solid grounding in physiology and in the experimental method, from the very beginning of medical education, will do much to remedy this regrettable state of affairs.

The desire to obtain the benefit of applied science must not make one forget that an exclusive interest in that which is immediately practical ends in a narrow rut of mediocrity. The most useful discoveries in medicine have been the consequence of disinterested research, not directed toward any definite application. Excessive preoccupation with practical results leads to shortsightedness and diminishes the fertility and scope of research and knowledge. There is no pure science; there is only science and its applications, which may be evident immediately or may take some time to develop. It is a common belief that many physiological facts are of only academic interest and have little or no practical value. This serious misconception should be corrected; every method of diagnosis and treatment now in use is the outcome of disinterested scientific research, and no one can say that knowledge that is apparently useless today will not have some application tomorrow that may be of great value.

Medical students can be taught only a small part of all that is known. The choice of what is most formative and useful will best be made by those with a wide knowledge of the science to be taught and with personal experience in its methods and techniques. Such competency can be achieved only after long and arduous work in the laboratory with complete dedication to the task.

A textbook of physiology for medical students and doctors must strike a balance between widely differing requirements. It must expound ideas taken from several sciences applied in physiology; it must explain the general principles of physiology and the functions of the different organs and systems, as well as their correlations; finally, it must integrate information obtained by the clinical and experimental study of altered function. A student can assimilate only a limited amount of knowledge in the time given to physiology in a well-balanced medical curriculum. This requires wise discrimination between the little that can be taught and the far wider field that must be left out. It is far better to have precise knowledge of fundamental principles than vague ideas about a great number of facts.

The facts of human physiology should be used in preference to physiological facts on other species for the illustration of general principles, thus satisfying the need for applied knowledge and at the same time linking it up with medicine and surgery. The tendency to separate them fostered by the inevitable separation between the clinic and the laboratory will thus be counteracted. It will be of great advantage for the education of the future physicians, and for the advancement of medicine and of physiology, if the development of as many contacts as possible is stimulated by the study of problems common to physiology and medicine.

Many problems can be solved only by animal experimentation, and most of the discoveries afterward usefully applied in man were first made by this method. Nevertheless research in human physiology has the advantage that its results can be applied immediately to man and that the subjects can give reasonable and conscious collaboration, which cannot be obtained from animals.

The study of pathologic conditions is also important, because knowledge of disease or altered function is useful for the understanding of normal function. Diseases can be considered as natural experiments, some of which have given valuable information not obtained from laboratory experiments. Clinical physiologists contribute daily to the advancement of physiology, and reciprocally a profound knowledge of physiology is indispensable for the making of original discoveries in clinical medicine and for efficiency in the practice of medicine. In the teaching of physiology, demonstrations in healthy subjects and a variety of clinical cases become more useful every day.

The student should never lose sight of the fact that it is only for didactic reasons that the different systems and organs are considered separately, as if they could be taken apart from the rest of the organism. This separation is completely artificial, as the whole organism is an indivisible anatomic and functional unit.

Physiology is in a constant process of evolution and advancement; a textbook should reflect this active life and not give the impression of a catalogue of completed and unchangeable statements. The reader's interest should be awakened by this progressive evolution, and the wish should be born in him to follow this continuous increase in knowledge and, if possible, to make personal contributions to it.

Science in general, and physiology in particular, can be taught properly only by intense, practical, and individual teaching—a fact recognized in all advanced centers of medical education. A book cannot be a substitute for this kind of teaching; it can be no more than a useful guide. The practical and rational study of physiology develops a scientific attitude of mind; namely, the capacity to find truth and to recognize error, and the habit of rigorous demonstration instead of dogmatic affirmation of imaginative statements that dazzle the simple-minded.

Physiology is nourished by experimental research, which is the permanent search for truth by adequate and precise methods. It is conscientious and continuous searching and researching so as to increase and perfect knowledge. In former times the only precise medical science was that of normal anatomy; later that of pathologic anatomy was added; at a still later date bacteriology was preeminent; today we are in the era of physiology, which has rejuvenated these sciences and given fresh vigor to them and to medicine in general.

In this book, statements will be supported by adequate proof as far as possible, so as to create the habit of critical appreciation. Unfortunately, for reasons of space it is not possible to do this in all cases, but it must be understood that these demonstrations exist and are available in more specialized books or in the original papers quoted in the bibliographic references. The only way of avoiding the intellectual vice of unreasoned acceptance of unsupported statements, when these are presented in an attractive form, is a practical and individual education in the scientific method. For such an education, no textbook can be a substitute.

Students do not always have sufficient training in the basic sciences of biology, physics, and chemistry. In other cases, their knowledge has been forgotten. It is therefore sometimes necessary to recall certain facts that ought to be familiar. Other indispensable information in anatomy, histology, embryology, biochemistry, and biophysics is omitted because it is considered as known or can be easily found in the appropriate textbooks. A few subjects of great importance are treated somewhat summarily, *e.g.*, nutrition and growth in childhood, and pregnancy and parturition. In the course of medical studies, these will be discussed in greater detail in pediatrics, obstetrics, etc. Certain information considered of secondary or mainly clinical interest, or debatable questions, or the description of methods, is included in smaller print. At the foot of the page reference is made to a few classic papers and to others of recent date that describe work of importance. At the end of each chapter there is a short list of books and papers which will help those who wish to make a deeper study of the subject in question.

Information included in this book has been obtained from sources in all countries and languages, and although sometimes the work done by the authors' group is given in greater detail, an effort has been made to keep this within reasonable bounds. The authors have been engaged for many years exclusively in teaching and research, and although they give their personal views and experience of the different problems, they have tried to render a balanced account covering the whole field of physiology. The book has been written in difficult circumstances, which have caused some otherwise avoidable defects. Also it is not easy to choose wisely and to present adequate knowledge in a field that is constantly developing over an ever wider area, extending to several allied sciences, and in which information must be kept up to date.

This book was written originally to satisfy the need of South American students. Later, French, Portuguese, and English editions were published. The authors have revised the text of each new edition; they have made many corrections, some of the chapters have been almost completely rewritten, and every effort has been made to keep the book up to date.

The authors will be grateful for constructive criticism and suggestions that may help to improve the book and make it a more efficient stimulus and guide for students, thus fulfilling the purpose with which it was written.

BERNARDO A. HOUSSAY

STANDARD UNITS

Mass: The unit is the gram (gram mass, gm.); 1 gm. is the mass of 1 ml. of water at 4°C.; it is the one-thousandth part of the international standard kilogram, made of platinum and iridium, which is kept in Paris.

Time: The unit is the second (sec.); 1 sec. is the sixtieth part of 1 min., which is the sixtieth part of 1 hr.; 1 hr. is the twenty-fourth part of the average solar day; *i.e.*, the average duration of the days in the solar year (86,400 sec.).

Length: The unit is the centimeter (cm.); 1 cm. is the hundredth part of the standard meter, which is the distance between two lines on a platinum bar kept in Paris; it is approximately one ten-millionth part of a quadrant of a meridian of the earth.

Velocity: The unit is the centimeter per second (cm./sec.).

Acceleration: The unit is the centimeter per second per second (cm./sec.^{2}).

Force: The CGS[1] unit is the dyne; 1 dyne is the force that, acting on a mass of 1 gm. for 1 sec., imparts to it a velocity of 1 cm./sec., *i.e.*, causes an acceleration of 1 cm./sec.^{2} The unit of gravity is the gram weight; 1 gram weight is the force with which the earth attracts 1 gram mass. It is equal to 981 (980.665) dynes and imparts to 1 gram mass an acceleration of 981 cm./sec.^{2} It varies in different places.

Work: The CGS unit is the erg; 1 erg is the force of 1 dyne acting through a distance of 1 cm.; *i.e.*, 1 erg = 1 dyne $\times$ 1 cm. = 23.8 $\times$ 10^{-9} cal. = 10.19 $\times$ 10^{-4} gm.-cm. One joule = 10^7 ergs = 0.23885 cal. The gravitational unit is the gram-centimeter; 1 gm.-cm. = 981 ergs = 23.43 $\times$ 10^{-6} cal., *i.e.*, 23.43 microcalories. The kilogram-meter (kg.-m.) is the most commonly used unit; it is the amount of work spent in raising a mass of 1 kg. to a height of 1 m. It is equal to 9.81 joules.

[1] Centimeter-gram-second, or metric, system of measurement.

Power: The CGS unit is the erg per second. One watt = 10^7 ergs/sec. = 1 joule per second. One kilowatt (kw.) = 10^3 watts = 10^{10} ergs/sec. The gravitational unit is the gram-centimeter per second. The metric horsepower is 75 kg.-m./sec. = 0.736 kw. The English horsepower (hp.) is 550 foot-pounds per second = 0.746 kw.

Heat: The unit is the calorie (cal.) or small calorie; 1 cal. is the amount of heat required to raise the temperature of 1 gm. of distilled water 1 degree centigrade (from 14.5 to 15.5°C.); it is equivalent to 4.185 $\times$ 10^{-7} erg = 0.4266 kg.-m. = 4.185 joules. One kilogram-calorie (kg.-cal.) or large calorie (Cal.) is 1,000 cal. = 426.6 kg.-m. = 4,1868 joules.

Temperature: The unit is the degree centigrade (°C.); 1 degree centigrade is the one-hundredth part of the difference in the temperature of melting ice and water boiling at a pressure of 760 mm. Hg. Zero in the centigrade scale corresponds to 273° (273.15°) in the scale of absolute temperature (T or K).

Mole is the molecular weight of a substance in grams = 6.02 $\times$ 10^{23} molecules (Avogadro's number). One millimole = 10^{-3} gram molecular weight. One micromole = 10^{-6} gram molecular weight.

Molar solution contains 1 mole or the molecular weight of the solute in grams in 1 liter of solution. It exerts an osmotic pressure of 22.4 atmospheres at 0°C., and its freezing point is —1.86°C. if the solute is not dissociated.

ABBREVIATIONS AND SYMBOLS FOR STANDARD UNITS

Prefixes: micro = 10^{-6}, *i.e.*, one-millionth
milli = 10^{-3}, *i.e.*, one-thousandth
centi = 10^{-2}, *i.e.*, one-hundredth
deci = 10^{-1}, *i.e.*, one-tenth
kilo = 10^{3}, *i.e.*, one thousand
mega = 10^{6}, *i.e.*, one million

Unit	Abbreviation
kilometer	km.
meter	m.
decimeter	dm.
centimeter	cm.
millimeter	mm.
micron (10^{-3} mm.)	μ
millimicron (10^{-6} mm.)	mμ
micromicron (10^{-9} mm.)	$\mu\mu$
Ångstrom unit (10^{-7} mm., 100$\mu\mu$)	Å
square meter	sq. m.
square centimeter	sq. cm.
degree (angle)	°
minute (angle)	′
second (angle)	″
hour	hr.
minute (time)	min.
second (time)	sec.
millisecond (0.001 sec.)	σ or msec.
microsecond (0.001 msec.)	μsec
kilogram-meter	kg.-m.
kilogram	kg.
gram	gm.
decigram	dg.
centigram	cg.
milligram	mg.
microgram (10^{-3} mg.)	γ or μg
liter	liter or l.
centiliter	cl.
milliliter (1,000,027 cc.)	ml.
cubic meter	cu. m.
cubic decimeter	cu. dm.
cubic centimeter	cc.
calorie	cal.
Calorie or kilogram-calorie	Cal. or kg.-cal.
degree Celsius (centigrade)	°C.
degree Fahrenheit	°F.
ampere	amp.
volt	volt or v.
millivolt (10^{-3} volt)	mv.
microvolt (10^{-6} volt)	μv.
cycles per second	c.p.s.
revolutions per minute	rev./min.

Contents

V. Metabolism and Nutrition

VI. Internal Secretions

VII. Reproduction

VIII. The Formation and Excretion of Urine

IX. The Nervous System

SECTION ONE

The Internal Environment and the Blood

CHAPTER 1

The Internal Environment and the Blood

THE INTERNAL ENVIRONMENT

The body fluids. Water constitutes about two-thirds of the body of an adult man. It can be said that the organism is made up of water in which micellae, molecules, and ions are dispersed. The physical and chemical reactions of colloidal and true solutions are therefore of fundamental importance in living organisms.

The body fluids are distributed in three compartments. One is within the cells, the *intracellular fluid;* the other two, outside the cells and forming the *extracellular fluid*, are the *interstitial fluid* and the *blood plasma*. Between these last two compartments the exchange of diffusible substances is easy and continuous, so they have a very similar content of water and salts.

Fluid blood is made up of a liquid, the blood plasma, in which cells (erythrocytes, leukocytes, and platelets) and minute particles (hemokonia or chylomicrons) are suspended. When blood coagulates, it first becomes solid and then a fluid oozes out, which is called the serum.

The internal environment. Multicellular organisms are surrounded by an external environment—air or water—but their cells live in a fluid environment which Claude Bernard in 1878 named the "*milieu intérieur*" (internal environment). This *milieu* is formed by the extracellular fluids, *i.e.*, by (*a*) the interstitial or tissue fluids, which bathe the cells and circulate slowly; (*b*) the lymph contained in the lymphatic vessels, which goes from the tissues to the blood; (*c*) the blood plasma, which circulates rapidly. The cerebrospinal fluid, the aqueous humor of the eye, and the fluid in the pleural and peritoneal cavities, the joints, and the synovial sheaths are particular forms of interstitial fluid.

The blood is the part of the *milieu intérieur* that circulates rapidly within a closed system of vessels. An outstanding feature of the blood is the constancy of its chemical composition and physical properties, thus assuring constant conditions for the functioning of the cells. The blood is being continuously renewed by incoming and outgoing cells and substances; nevertheless the variations in its composition fluctuate within a narrow margin and are rapidly corrected. The functions of the organism are regulated so as to maintain the stability of the internal environment, a physiological fact of great importance which Claude Bernard first pointed out and called the "fixity of the *milieu intérieur*." The concept was developed by Cannon, who coined the term "homeostasis" to signify those steady states maintained by coordinated complex physiologic reactions.

Therefore, on the one hand the blood assures constant conditions to the cells, and on the other the organism maintains the chemical, physical, and morphologic constancy of the blood. According to Claude Bernard, the fixity of the internal environment is the necessary condition for the free and independent life of higher organisms.

THE BLOOD

Blood consists of a fluid in which there are free cells—the red cells, or erythrocytes; the white cells, or leukocytes; and the platelets. It is also a dispersion of micellae, molecules, and ions in water, and therefore has the properties of colloidal and true solutions.

The relative volumes of erythrocytes and plasma. The volume of cells in 100 cc. of

tion) with the plasma above. The specific gravity of the blood is conditioned mainly by the number of red blood cells; it is greater in man, 1.059 (1.052 to 1.063), than in woman, 1.056 (1.050 to 1.058). The specific gravity of plasma and of serum is principally due to the

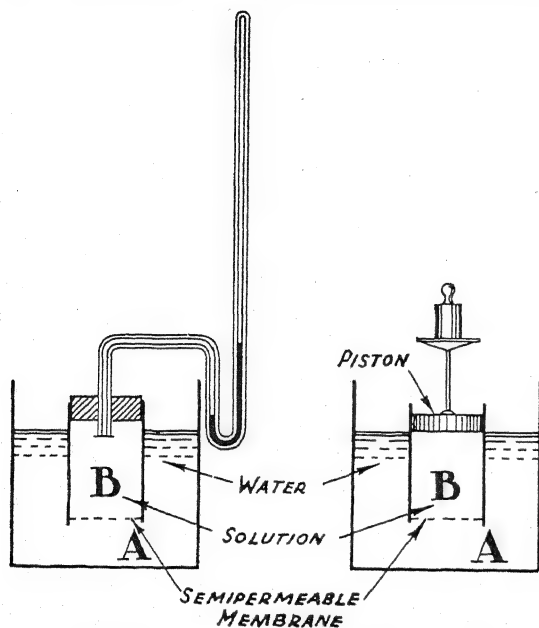



FIG. 2. Demonstration of osmotic pressure.

protein concentration and is therefore used to measure approximately the protein content.[1] As the method is simple and rapid, it is frequently used to follow variations in plasma-protein concentration in cases of shock, burns, etc.

The specific gravity of whole blood increases and decreases with the erythrocyte concentration. It is possible to calculate the hemoglobin concentration of blood by measuring the specific gravities of whole blood and of plasma.[2]

The specific gravity of the blood plasma or serum can be determined by weighing a standard quantity in a pyknometer. Another method consists in comparing the specific gravity of a drop of blood with that of solutions of known specific gravity. Several techniques have been proposed; that of Phillips and Van Slyke[3] has the advantage of simplicity. A drop of blood is introduced into each of a series of tubes containing different concentrations of copper sulfate of known specific gravity. The drop is surrounded by a thin coat of coagulated protein, and it falls if its specific gravity is higher than that of the solution but rises if it is lower. If the specific gravity of the blood is the same as that of the solution, the drop remains in equilibrium after 10 to 15 sec. Three or four drops of blood are sufficient, and the whole procedure can be completed in 1 min.

[1] VAN SLYKE, D. D., *et al.*, *J. Biol. Chem.*, **183**, 331, 1950.

[2] *Ibid.*, p. 349.

[3] PHILLIPS, R. A., D. D. VAN SLYKE, *et al.*, *J. Biol. Chem.*, **183**, 305, 1950.

Viscosity. A fluid in movement has the property of viscosity because of internal friction between its constituent particles. The viscosity of whole blood is strictly dependent on the red cell count. In man the relative viscosity is on an average 4.7 (4.3 to 5.3) and in woman 4.4 (3.9 to 4.9), taking that of distilled water as unity, and measuring it in the Hess viscosimeter.[1] The viscosity of plasma or serum is lower than that of whole blood; measured in Ostwald's viscosimeter, it is 1.6 to 2.2 for serum, and about 20 per cent higher for plasma. The viscosity of plasma is in direct proportion to the total protein concentration, and especially to the concentration of serum globulin, which has large, long, and asymmetric molecules. In hypothyroidism (thyroid insufficiency) there is an increase in blood globulins, and therefore in the viscosity of the plasma; while in hyperthyroidism (excess thyroid function) plasma globulins and viscosity diminish.

Osmotic pressure. This is pressure developed by water[2] when it passes through a semipermeable membrane from a dilute solution to a concentrated solution (Fig. 2). A semipermeable membrane[3] is permeable only to water and not to the substances in the solution. Osmotic pressure can be demonstrated in the following way (Fig. 2): A solution of sugar in water is placed in a flask *B*, the bottom of which is a semipermeable membrane; the top is hermetically closed by a piston or by a stopper pierced by the tube of a manometer. The flask is submerged in a larger receptacle *A* full of water. Molecules of water will pass through the membrane in both directions, but in much greater number

[1] HESS, W. R., *Deutsch. f. klin. Med.*, **94**, 404, 1908.

[2] Only aqueous solutions are mentioned, as these are the only ones of importance in living organisms.

[3] Pfeffer used a copper ferrocyanide membrane, but others are also used, although most of them are not strictly semipermeable, as other molecules besides those of water pass through them.

from *A* to *B* than from *B* to *A*, and the piston will rise in *B* or the manometer will mark a rise in pressure. If an adequate weight is placed on the piston, the same number of molecules of water will be forced through the membrane from *B* to *A* as that passing from *A* to *B*, so the piston will remain stationary. The weight needed to keep the piston from rising is the equivalent of the hydraulic pressure developed by the osmotic pressure and therefore serves to measure it.

The substance dissolved in a solution occupies space within the solvent; therefore the concentration of water in *A* (pure water) is greater than in *B* (solution); more molecules of water will therefore bombard the semipermeable membrane on the side of *A* than on that of *B*, and more will pass through it from *A* to *B* than from *B* to *A*.

The osmotic pressure developed by a solution is equivalent to the pressure that the dissolved substance would develop if it were a gas at the same pressure and temperature. One gram molecule of any gas at 760 mm. Hg and 0°C. has a volume of 22.4 liters; compressed to a volume of 1 liter, it develops a pressure of 22.4 atmospheres at 0°C.

The gas laws can therefore be applied to solutions. Consequently, (*a*) solutions with the same osmotic pressure at the same temperature contain the same number of molecules in unit volume (Avogadro's law); (*b*) at a constant temperature the osmotic pressure is proportional to the molecular concentration (the Boyle-Mariotte law); (*c*) with a constant molecular concentration the osmotic pressure is proportional to the absolute temperature (Gay-Lussac's law); (*d*) if several substances are dissolved in the same solvent, each one develops osmotic pressure as if it were the only substance dissolved; therefore the total osmotic pressure is the sum of the partial osmotic pressures of the substances in solution (Dalton's law). These laws are valid for dilute solutions; in concentrated solutions there are other factors that limit their applicability.

Electrolytes in solution are partially dissociated into their constituent ions; the osmotic pressure of these solutions is that developed by the concentration of all the particles, molecules and ions. Thus the osmotic pressure of a molar solution of NaCl (58.5 gm. per liter) is not the same as that of a molar solution of glucose (180 gm. per liter) but almost twice as much (1.94), because the salt is almost completely dissociated into its two ions.

Osmotic pressure is usually measured by indirect methods, because they are simpler and less open to error than direct ones. Indirect methods are based on the fact that several properties of solutions are conditioned by the concentration in particles (molecules, ions, and micellae). These properties are the following: osmotic pressure, freezing point, boiling point, and evaporation pressure. The freezing point of a solution is lower than that of the pure solvent, and the depression of the freezing point is proportional to the concentration of particles in the solution. By measuring this depression, therefore, the osmotic pressure can be calculated. A molar solution (1 gram molecule per liter) freezes at −1.86°C. and has an osmotic pressure of 22.4 atmospheres at NTP (normal temperature and pressure, *i.e.*, 0°C. and 760 mm. Hg).

Solutions of the same osmotic pressure are called isotonic. Of two solutions, the one with the higher osmotic pressure is hypertonic with respect to the one with the lower osmotic pressure, and reciprocally this second one is hypotonic with respect to the first.

The osmotic pressure of whole blood is approximately the same as that of plasma or of serum. The lower marine animals have more or less the same osmotic pressure as the sea water in which they live; they are known as "poikilosmotic" because their osmotic pressure is dependent on that of the external environment. More developed species have a constant osmotic pressure; they are called "homoiosmotic" because the pressure of their blood plasma is maintained at a constant level. Human plasma freezes at −0.56°C. (−0.54 to −0.59°C.), which corresponds to a 0.3 molar solution, and an osmotic pressure of 6.7 atmospheres. Seventy-five per cent of this pressure is due to NaCl. A 0.9 per cent NaCl solution is isotonic with plasma (has the same osmotic pressure); this is known as normal saline or physiologic solution.

The osmotic pressure developed by the plasma proteins is quantitatively much less than that of other plasmatic substances, but it is of great importance (Starling). Water and crystalloids diffuse easily through the membranes of the capillaries, so they are in approximately the same concentration in the interstitial fluid as in the blood plasma. The plasma proteins, on the contrary, do not diffuse readily through the capillary membrane and, as they remain within the blood vessels, they retain water. To make

water leave the capillaries, it is necessary to apply, within the blood vessels, a hydrostatic pressure higher than the osmotic pressure of the plasma proteins.

The osmotic pressure of the plasma proteins can be measured by placing plasma, or the plasma proteins, in a small osmometer with a cellophane membrane submerged in a fluid of the same saline concentration as plasma, *e.g.*, Ringer's or Locke's solutions, or in plasma ultrafiltrate. In this case the pressure developed (colloidal osmotic or oncotic pressure) is determined by two factors: (*a*) the true osmotic pressure, *i.e.*, pressure dependent on the molecular concentration of protein; (*b*) pressure developed by hydration and consequent swelling of the protein micellae (imbibition).

Table 1. Chemical Composition of Blood, Plasma, and Erythrocytes

Constituents	*Blood*	*Plasma*	*Erythrocytes*
	Gm./100 cc.		
Water	78.0 (77–85)	90.7	66.0
Total solids	22.0 (18–23)	9.3	34.0
Organic substances	21.2	8.5	33.0
Salts	0.8 (0.6–1)	0.93	0.7
Total protein	18.5	7.0	30.0
Serum albumin	2.5	4.2	
Serum globulin	1.38	2.6	
Fibrinogen	0.25	0.3	
Hemoglobin	15.0 (13–17)	....	34.0
Total nitrogen	3.3	1.2	5.3
	Mg./100 cc.		
Nonprotein N	33.0	25.0	44.0
Urea N	12.0	15.0	11.0
Amino-acid N	5.6	4.5	7.4
Undetermined N	13.0	3.0	25.0
Urea	20.0–35.0	26.0	2.0
Uric acid	2.0	3.0	2.0
Creatinine	1.1	3.3	0.7
Creatine	0.4	0.42	3.1
Ammonia	0.25		
Indican	....	0.6	
Phenol	1.6	1.7	1.5
Bilirubin	....	0.6	
Glucose	70.0	80.0	65.0
Lactic acid	6.0	8.0	5.0
Fatty acids	360.0	370.0	340.0
Lecithin	300.0	200.0	400.0
Cholesterol	200.0	180.0	200.0
Ketonic bodies	2.0		
Total phosphorus	45.0	10.0	75.0
Acid-soluble P	30.0	25.0	50.0
Inorganic P	5.0	3.5	6.0
Ester P	24.0	.....	24.0
Lipid P	13.0	7.0	18.0

The plasma proteins develop an osmotic pressure of 25 to 35 mm. Hg. The osmotic pressure of serum albumin is much greater than that of serum globulin; therefore its capacity to retain water is also greater. This difference is due to the smaller size of the molecule of serum albumin, so that in the same weight there are more molecules (particles) of serum albumin than of serum globulin. Also the oncotic pressure of serum albumin is greater. Serum albumin accounts for 80 per cent of the total osmotic pressure developed by the plasma proteins. One gram of serum albumin retains 18 cc. of water at 22 mm. Hg, 25°C., and pH 7.4, and is equivalent in this respect to 20 cc. of citrated plasma. A solution of 25 gm. of serum albumin in 100 cc. of water can retain within the blood vessels as much water as 500 cc. of citrated plasma.[1] This fact has led to the use of plasma or serum albumin solutions in the treatment of shock produced by wounds and burns, with the object of increasing the blood volume by attracting and retaining water within the blood vessels. This has been one of the major advances in therapeutics of recent years.

When the concentration of plasma proteins or of serum albumin diminishes, water and salts easily diffuse from the capillaries because the blood pressure is much higher than the osmotic pressure of the plasma proteins. In hypoproteinemia and hypoalbuminemia the interstitial fluid increases and edema results.

Chemical Composition

The principal constituents of blood and their average concentration are given in Tables 1 and 2.

INORGANIC CONSTITUENTS

Plasma contains 90 per cent water and the erythrocytes only 65 per cent. There are also differences in the mineral composition of plasma and cells. Thus extracellular fluid, *i.e.*, plasma and interstitial fluid, has a preponderance of Cl, Na, and Ca; human erythrocytes, on the

[1] Scatchard, G., A. C. Batchelder, and A. Brown, *J. Clin. Investigation,* **23**, 458, 1944.

other hand, contain almost all the iron there is in the blood, much more K than plasma, and a little more Mg. Nevertheless the small quantities of plasmatic K and Mg are of great physiologic significance. These elements, together with Na and Ca, maintain an ionic equilibrium which is of fundamental importance in the life and function of cells.[1]

Table 2. Inorganic Constituents of Plasma

	Constituents, mg./100 ml.										*Total base cc. 0.1 N Na(OH)*	*Alkaline reserve, CO_2 cc./100 ml.*
	Na	*K*	*Ca*	*Mg*	*Cl*	*NaCl*	*P*	*Fe*	*Cu*	*I*		
Plasma......	330 (320–360)	17* (14–20)	10 (9–11.5)	3 (2.5–3.5)	365 (340–370)	600 (580–620)	3.5 (2.6–5.4)	0.12 (0.05–0.18)	0.10 (0.08–0.16)	0.010 (0.008–0.012)	155	60 (55–75)
Cells......	45	410	1.0	4	185	...	...	100	....	0.005		
Blood........	190	200	5.6	3	280	480	...	50	0.14	0.008		
mEq./l. of plasma.....	142	4.1	5.0	3	103	...	1.1	....	....	......	155	27 mm. (20–31)

*When separation of the erythrocytes has been delayed, K in plasma rises to 20 mg. (18 to 22 mg.) per 100 ml.

Seventy-five per cent of the molecules in blood plasma are electrolytes, and of the latter three-quarters are NaCl. Na constitutes 92 per cent of the base in plasma, while K makes up almost all the base in the human erythrocyte.

Phosphorus in the blood is found in four different kinds of combination: *inorganic* (orthophosphate), and three *organic*, which are *ester phosphorus* (glycerophosphates, hexosephosphates), *lipid phosphorus* (phosphatids), and a very small amount of *nucleic acid phosphorus*. The inorganic P is less than one-tenth of the total P.

The concentration of the inorganic constituents is frequently expressed in milliequivalents per liter (mEq./liter), because the elements combine with each other in equivalent quantities. The chemical equivalent of an element in a solution is calculated by dividing the concentration in grams per liter by the atomic weight and multiplying by the chemical valence. Thus, the concentration of Ca in blood plasma is 100 mg. per liter, its atomic weight is 40.07, and it is divalent. Therefore $\frac{0.100 \times 2}{40.07} =$ 0.0049 equivalents per liter, or $0.0049 \times 1{,}000 = 4.9$ mEq./liter. There is 3.35 gm. per liter Na in blood plasma, its atomic weight is 23, and it is monovalent; therefore $\frac{3.35 \times 1}{23} = 0.1456$ equivalents per liter or 145.6 mEq./liter of Na in plasma. Concentration in milliequivalents per liter can be obtained by multiplying the concentration, in milligrams per 100 ml., by the following figures: Na, 0.431; K, 0.25; Ca, 0.5; P, 0.324; Cl, 0.284; CO_2 (volume in cubic centimeters per 100 ml.), 0.45.

The physiologic significance and variations of the inorganic constituents of blood plasma will be considered in Chap. 45, when dealing with mineral metabolism.

PLASMA PROTEINS

Amount and fractions. The total concentration of plasma proteins is about 7.0 gm. per cent (6 to 8 gm. per cent). When salts[1] are added the proteins are precipitated in the following order: fibrinogen, euglobulins, pseudoglobulins, and serum albumin. Proteins with large molecules are less stable than those with smaller molecules. Thus fibrinogen has a molecular weight of 500,000, that of the various globulins is about 150,000, and that of serum albumin is

[1] There is a mistaken idea that the mineral content of blood plasma is similar to that of sea water (Bunge, 1889; Quinton, 1897; Macallum, 1903). Blood plasma has a lower total mineral concentration than sea water, and relatively more K and less Mg and sulfate (MACALLUM, *Physiol. Rev.*, **6**, 316, 1926).

[1] Ammonium sulfate or sodium sulfate (Howe, 1921), in gradually increasing concentrations, is used. Different concentrations of alcohol at low temperature (5°C.) have been used with great advantage, because at this temperature alcohol does not alter the proteins (Cohn).

69,000 (Fig. 3). Serum albumin (the highest concentration) is 4 to 5.6 gm. per cent; globulins, 1.5 to 3 gm. per cent; and fibrinogen, 0.20 to 0.4 gm. per cent. The albumin-globulin ratio varies from 1.5 to 2. With the albumin a small quantity of glycoprotein[1] is precipitated (serum mucoid or globomucoid).

Fibrinogen is responsible for the coagulation of the blood when it changes from a hydrosol to a gel, called fibrin (see Chap. 8); it is unstable, and when plasma is heated it coagulates at 56 to 60°C., before the other proteins. Globulins are the most viscous of all the plasma proteins.

The antibodies, *i.e.*, the antitoxin activity of antitoxic serums, are found in the euglobulins; when these are precipitated and redissolved the solution is called concentrated antitoxic serum or antitoxin.[2] Serum albumin can be obtained in crystallized form; it is the plasma protein with the highest osmotic pressure.

The protein content is estimated by precipitating the proteins with salts, or coagulating them by heat or acids, and then weighing the precipitate or determining the nitrogen content. The plasma protein content can be calculated by colorimetric estimation of some amino acids. The specific gravity and the refractive index of plasma have also been used for determining the protein content. These methods are simple and quick, and for this reason are much in use for clinical purposes.

In an electric field, proteins in solution migrate toward one of the poles; this is known as electrophoresis. The velocity of migration depends on the electric charge of each particle. At an adequate pH the different proteins migrate at different velocities. By means of a photographic method developed by Tiselius (1937), it is possible to register the velocity and relative concentration of the different proteins. The order of velocity for the plasma proteins is the following: serum albumin, α globulins (α_1, α_2), β globulins (β_1, β_2), fibrinogen, and γ globulin (the latter having the lowest velocity). This method gives a more detailed analysis of the plasma proteins than does precipitation.

The relative amounts of the different proteins are the following: serum albumin, 55 per cent; α globulin, 13 per cent; β globulin, 14 per cent; γ globulin, 11 per cent; fibrinogen, 7 per cent. During the Second World War, fractionation of the plasma proteins for therapeutic use was greatly developed. Serum albumin is used in transfusion to restore the blood volume; fibrinogen, to accelerate blood clotting in wounds; γ globulin, in the prevention and treat-

Fig. 3. Relative dimensions of several proteins. (*Cohn, E. J., Ann. Int. Med., vol. 26, p. 342, 1947.*)

[1] Mannose, galactose, and hexosamine form part of its molecule.

[2] This preparation does not contain pure antitoxin; part of the inactive substances can be eliminated by partially digesting the serum; antitoxin resists this operation and can be precipitated afterward (see Chap. 9).

ment of measles. E. Cohn[1] has directed a committee which has made numerous and important studies on the chemistry and physicochemistry of blood plasma and on its therapeutic and immunologic properties.

Functions of the plasma proteins. Plasma proteins are of great physiologic importance. A marked decrease in their concentration (hypoproteinemia) is accompanied by lesions in several organs, delay in the healing of wounds, and edema. The edema is due to accumulation of interstitial fluid, owing to the fact that water escapes from the capillaries because of the low osmotic pressure of the blood plasma. The principal functions of the plasma proteins are the following:

1. *Nutrition.* Total plasma or serum albumin given in intravenous injection is used in protein metabolism, and in the formation of plasma proteins and hemoglobin.
2. *Blood clotting.* This is due to the transformation of fibrinogen into a gel—fibrin—by the action of thrombin (see Chap. 8).
3. *Blood viscosity.* The proteins with larger and more asymmetric molecules (globulins) play a more important part than serum albumin in giving the blood a certain viscosity, which is an important factor in conditioning the work of the heart and in maintaining the blood pressure.
4. *Osmotic pressure.* Plasma proteins, owing to their osmotic pressure, are important for the retention of water within the capillaries (see page 7).
5. *Suspension stability of the erythrocytes.* This is dependent principally on fibrinogen, less on globulin, and much less on albumin concentration (see Chap. 3).
6. *Immunity.* The active principles in immune serums are precipitated by salts together with the euglobulins, more specifically with the γ globulins and partly with β globulin separated by electrophoresis.
7. *Acid-base equilibrium.* A protein behaves as a weak acid when in solution in a medium that is alkaline with respect to the isoelectric point of the protein; in an acid medium it behaves as a weak alkali.

The formation of plasma proteins. Plasma proteins are being continuously formed and destroyed. They do not constitute an inert store, but are in a constantly changing dynamic equilibrium. There is a small reserve of plasma proteins in the tissue, and they can be rapidly regenerated.

Plasmapheresis has been used in the study of the regeneration of plasma proteins. The animal is bled and the erythrocytes, suspended in saline solution, are reinjected, thus removing the plasma proteins. When the plasma protein concentration falls below 1 or 2 per cent, severe shock results. By provoking a less marked depletion, practicing plasmapheresis once a week, it is possible to demonstrate that several foodstuffs have a qualitative and quantitative influence on the formation of plasma proteins (Whipple).[1] Ingestion of amino acids increases the formation of plasma proteins. An inadequate diet can be the cause of hypoproteinemia, and in this case there is a deficient regeneration of plasma proteins after hemorrhage. Infection has a similar effect.

Plasma proteins and hemoglobin can be formed when the following substances are injected intravenously: plasma; serum; the product of digestion of plasma or serum; the product of advanced digestion of hemoglobin or of casein; or the amino acids indispensable for nitrogen equilibrium (Elman, Whipple).

Fibrinogen is rapidly regenerated; it is almost completely restored in 6 to 24 hr., but not if the liver has been removed, as it is formed in this organ.[2] There is a small extrahepatic store of fibrinogen. The regeneration of other plasma proteins is considerably delayed after hepatectomy, so probably the liver plays an important part in their formation.

There is a lower concentration of proteins and of serum albumin in children than in adults. During pregnancy albumin decreases, and fibrinogen and in a lesser degree globulins increase.

In many acute infectious diseases, fibrinogen increases; in several chronic infectious diseases and in the course of immunization, globulins increase. In nephrosis albumin is lost in the urine, so its concen-

[1] COHN, E., *et al.*, *J. Clin. Investigation*, **23**, 4, 1944 (23 papers); *Ann. Int. Med.*. **26**. 341. 1947.

[1] MADDEN, S. C., and G. H. WHIPPLE, *Physiol. Rev.*, **20**, 194, 1940. WHIPPLE, G. H., and S. C. MADDEN, *Medicine*, **23**, 215, 1944; MADDEN, S. C., *et al.*, *J. Exper. Med.*, **82**, 181, 1945.

[2] MCMASTER, P. D., and D. R. DRURY, *Proc. Soc Exper. Biol. & Med.*, **26**, 490, 1928; JONES, T. B., and H. P. SMITH, *Am. J. Physiol.*, **94**, 144, 1930.

tration in plasma diminishes and consequently edema is produced. The thyroid gland has a marked influence on the plasma globulins; α globulin increases in hypothyroidism (thyroid insufficiency) and decreases in hyperthyroidism (excess thyroid function).

NONPROTEIN NITROGEN

After the plasma proteins have been completely removed, the plasma contains only crystalloid nitrogenous compounds. The nitrogen of these substances is known as nonprotein nitrogen; its concentration is from 25 to 35 mg. per 100 cc.[1] (see Table 1).

The plasma proteins are usually removed by precipitation, but they can also be separated by ultrafiltration. Urea is the principal nonprotein nitrogenous substance; about one-half of the total nonprotein nitrogen is in the form of urea; then in decreasing order of concentration come the amino acids, uric acid, creatinine (usually determined as creatine),[2] and finally an important fraction called the undetermined nitrogen (Table 1).

Three factors have an influence on the nonprotein nitrogen: (*a*) *deficient renal excretion;* (*b*) *excess production;* (*c*) *increased fixation in the tissues.* In renal insufficiency nonprotein nitrogen accumulates in the blood, because of deficient excretion; the determination of its concentration in blood, or of urea or creatinine concentration, is used in the diagnosis and prognosis of renal disease. Nonprotein nitrogen also increases in the following cases: when there is severe dehydration, shock, acute adrenal insufficiency, hemorrhage into the digestive tract; after the ingestion of hemoglobin; and in some acute infectious diseases. In many of these cases there seems to be an excess production of nonprotein nitrogenous compounds. A decrease in nonprotein nitrogen concentration is observed only occasionally at the end of pregnancy, and more markedly a few hours after injecting the growth-promoting extract of the anterior hypophysis. In both these cases it is supposed that this decrease is due to increased fixation of nitrogenous substances by growing tissues.

[1] Folin, O., *Physiol. Rev.*, **2**, 460, 1922.

[2] Only about 80 per cent of what is estimated colorimetrically in the plasma as creatine is really creatine, and from 30 to 40 per cent of what is estimated in the erythrocytes (Beard).

OTHER CONSTITUENTS

Blood contains other inorganic substances beside those already mentioned, and several organic substances, such as glucose, lipids, cholesterol, enzymes, hormones (internal secretions), and antibodies (immune substances). Blood also has several gases in solution or in combination. All these constituents will be studied in following chapters. The acid-base equilibrium of the blood will be considered in the discussion of respiration (see Chap. 29).

BIBLIOGRAPHY

Hematology

Downey, H., "Handbook of Hematology," 4 vols., Hoeber, New York, 1938.

Whitby, L. E. H., and C. J. C. Britton, "Disorders of the Blood," 7th ed., J. & A. Churchill, London, 1954.

Wintrobe, M. M., "Clinical Hematology," 3d ed., Lea & Febiger, Philadelphia, 1951.

Chemical Composition

Diggs, L. W., Normal Human Values Obtained by Laboratory Procedures, in O. Glasser's "Medical Physics," vol. 2, Year Book Publishers, Inc., Chicago, 1950.

Krebs, H. A., Chemical Composition of Blood Plasma and Serum, *Am. Rev. Biochem.*, **19**, 409, 1950.

Myers, V. C., and E. Muntwyler, *Physiol. Rev.*, **20**, 1, 1940.

Peters, J. P., and D. D. Van Slyke, "Quantitative Clinical Chemistry," 2d ed., Williams & Wilkins, Baltimore, 1946.

Sunderman, F. W., and F. Boerner, "Normal Values in Clinical Medicine," Saunders, Philadelphia, 1950.

Plasma Proteins

Brand, E., Amino Acids in Proteins, *Ann. New York Acad. Sc.*, **47**, 187, 1946.

Cohn, E., *et al.*, *J. Clin. Investigation*, **23**, 4, 1944 (23 papers).

Edsall, J. T., *Advances in Protein Chemistry*, **3**, 384, 1947.

Madden, S. C., *et al.*, *J. Exper. Med.*, **82**, 181, 1945 (regeneration of plasma proteins).

Madden, S. C., and G. H. Whipple, *Physiol. Rev.*, **20**, 194, 1947.

Metcoff, J., and J. J. Stare, *New England J. Med.*, **236**, 26, 1947.

Muntwyler, E., *J. Lab. & Clin. Med.*, **30**, 526, 1945.

Tullis J. L., Mem. Univers. Labor Physic. Chem., Harvard University, 1953.

Wuhrmann, F., and C. Wunderly, "Die Bluteiweisskörper des Menschen," Schwabe, Basel, 1947.

CHAPTER 2

The Blood Volume

THE RED CELL COUNT and hematocrit readings give the concentration of erythrocytes and the percentage of plasma, but they give no information as to the total quantity of red cells or plasma. Knowledge of the total amount of blood, plasma, and erythrocytes is of great importance for the understanding of the physiology and pathology of the blood, the circulation, and respiration.

METHODS FOR THE ESTIMATION OF BLOOD VOLUME

The total amount of blood, plasma, and erythrocytes can be measured by direct and indirect methods. The data obtained could be expressed in weight, but it is usual to give them in volume and to speak of total blood volume, plasma volume, and erythrocyte volume.

DIRECT METHOD

Welcker's method (1854) is still used as a standard of comparison when studying new methods. It consists in the extraction of all the blood by collecting that which flows from an artery and that obtained by washing out the tissues. The procedure is as follows: A small quantity of arterial blood is collected and diluted to 1 per cent in saline. All the blood flowing from a large artery is then collected, and when the flow ceases the vascular system is washed out by intravenous injection of saline or Ringer's solution. The animal is then cut up into small pieces, which are extracted with saline. The blood and the fluids obtained by washing and extraction are well mixed and diluted until the color is the same as that of the first sample of 1 per cent arterial blood. The total quantity of fluid so obtained divided by 100 corresponds to the total blood volume of the animal.

Bischoff (1855) applied the method in man, using the bodies of two criminals who had been beheaded. The total amount of blood thus determined was found to be equal to 7.7 per cent of the body weight. This method has limited applications, as even in experimental animals it can be used only when survival is not indispensable. The determination is also subject to certain causes of error: (*a*) the pigment in muscle is extracted; (*b*) cloudiness of the fluid makes the colorimetric comparison difficult; (*c*) some erythrocytes can remain in the clots or in the small capillaries.

INDIRECT METHODS

Usually only the total plasma volume or total red cell volume is measured; then the plasma-cell ratio is determined by means of the hematocrit and thus the total blood volume can be approximately established.

There are two types of indirect methods: (*a*) a certain quantity of fluid is injected intravenously and the dilution suffered by a normal constituent of the blood, *e.g.*, erythrocytes or proteins, is measured; (*b*) a certain quantity of a foreign substance is introduced into the circulation, and once it has been distributed throughout the blood, its concentration is measured. Methods of the second type are the only ones now in use.

Carbon monoxide method.[1] This method is based on the fact that CO makes a stable combination with the hemoglobin in the erythrocytes. The oxygen capacity of the subject's blood is determined, *i.e.*, the

[1] First used by Grehant and Quinquaud in animals (1882) and by Haldane and Lorrain-Smith in man (1900). It has been considerably improved by Chang and Harrop (*J. Clin. Investigation*, **5**, 393, 1928). CO is no longer determined colorimetrically, but by the more accurate van Slyke manometric method.

maximum quantity of oxygen absorbed by 100 cc. of blood. The oxygen capacity is the same as the CO capacity. The subject then breathes into a bag which contains a known volume of CO mixed with oxygen. Since CO is toxic, a small quantity should be given so that only a small percentage (about 15 per cent) of the hemoglobin is transformed into carboxyhemoglobin. The concentration of CO in the blood is determined, and the blood volume is calculated. For example, the inhalation of 120 cc. CO has saturated 15 per cent of the hemoglobin in the blood, which has a maximum oxygen capacity (CO capacity) of 20 cc. per cent; therefore $(120 \times 100)/15 = 800$ cc. of gas would be needed to saturate all the hemoglobin in the circulation. As the $O_2(CO)$ capacity is 20 per cent, the total blood volume will be $(800 \times 100)/20 = 4{,}000$ cc. Fixation of CO by hemoglobin in the muscle fibers (myohemoglobin) is a cause of error in this method. The oxygen capacity of the blood is diminished by the inhalation of CO; therefore this method should not be used in cases of anemia, heart insufficiency, or other conditions in which the respiratory function of the blood is impaired.

Radioactive iron method. Radioactive iron is given to a donor belonging to group O (see page 59) so that it will be found in the erythrocytes. A sample of blood is drawn and its radioactivity is measured. A known quantity of this radioactive blood is injected into the blood stream of the subject in which the blood volume is to be determined. The dilution of the radioactivity in the recipient's blood is measured, and the total blood volume can be calculated. This is a very accurate method, and its results are more reliable than those of the dye method.

Dye method. This is the method most frequently used. A measured amount of a colloid dye is injected intravenously, and by determining its dilution, the blood volume can be calculated. The dye must fulfill the following requisites: (*a*) it must be innocuous; (*b*) it must diffuse very slowly out of the blood vessels; (*c*) it must not stain the erythrocytes or other tissue cells; (*d*) it must mix uniformly throughout all the plasma in the body; (*e*) it must be possible to determine its concentration in plasma accurately, even when there is turbidity or hemolysis. Many dyes fulfill these conditions; those most used have been vital red, congo red, trypan red, and trypan blue, but now T 1824 (sometimes called Evans' blue) is preferred. The concentration in plasma is measured by spectrophotometry with a photoelectric colorimeter. The technique of this method has been studied and perfected in its details in several animal species, and it has been applied to man in health and disease.[1] The procedure is as follows: A measured quantity of the dye is injected into a vein of the arm. A 7-min. period is allowed for a thorough mixing of the dye in the plasma, after which a sample of blood is drawn from a vein of the other arm. The concentration of dye in the plasma is measured, and the total plasma volume is calculated (Gregersen, Gibson, and Evans). The plasma-cell ratio is measured with the hematocrit (see Chap. 1), and the total blood volume is calculated.

The distribution of the dye is dependent on the velocity of the circulation. This method therefore measures only the blood which is actively circulating, not that which is stored.

THE NORMAL BLOOD VOLUME

The blood volume can be measured in the basal state or in different physiologic conditions, *e.g.*, exercise. To measure it in the basal state the subject must be fasting, lying down and completely at rest both mentally and physically. The basal blood volume is fairly constant in the same subject, even when measured on different days.

The results obtained with the different methods vary to some extent. The carbon monoxide method gives a basal blood volume of 7.1 to 7.5 per cent of the body weight. The dye method gives higher figures, from 8 to 10 per cent of the body weight. The greater values probably exceed the true blood volume.

The blood volume is proportional to the body weight; it is not a linear function of weight, but varies in relation to $W^{0.72}$ (W = body weight). A child therefore has a greater blood volume per kilogram of body weight than an adult, and a small animal than a large one. After puberty the total blood, plasma, and cell volumes are greater in man than in woman. Gibson and Evans, using the dye method, found an average of 7.7 liters per 100 kg. in man and 6.6 liters per 100 kg. in woman; and in weight 8.2 kg. per 100 kg. and 7 kg. per 100 kg.

The blood volume is more closely proportional to body surface than to body weight.

[1] GREGERSEN, M. J., *J. Lab. & Clin. Med.*, 29, 1266, 1944.

Gibson and Evans give an average of 2,923 cc. per sq. m. in man and 2,523 cc. per sq. m. in woman.[1]

Plasma volume is more constant than cell volume and varies little in the same subject. The average per kilogram is 43 cc. (38 to 58 cc.) in man and 41 cc. in woman; per square meter of body surface it is 1,700 cc. in man and 1,600 cc. in woman.

Distribution of blood in the body. Bazett has calculated that in an adult with a total blood volume of 5.2 liters, this would be distributed as follows: 250 cc. in the heart; 1,300 cc. in the pulmonary circuit; 550 cc. in the arteries, 300 cc. in the capillaries, and 2,250 cc. in the veins (*i.e.*, 3,100 cc. in the major circuit—arteries, capillaries, and veins). The remaining 550 cc. is probably to be found in blood reservoirs (liver, spleen).

VARIATIONS AND REGULATION OF THE BLOOD VOLUME

Changes can take place in the volume of the whole blood or in that of only one of the blood constituents. The following variations can be observed:

1. Total blood volume can be transitorily diminished by hemorrhage or increased by blood transfusion.
2. Erythrocyte volume can diminish, as in anemia, or increase, as in polycythemia.
3. Leukocyte volume is increased in leukocytosis and diminished in leukopenia.
4. Plasma volume can diminish, as is observed in cases of severe burns in which there is transudation of plasma on the burned surfaces. It is increased by transfusion of plasma, and for a short time after intravenous injection of saline solution.
5. Plasma water diminishes in anhydremia (lack of water in the blood) consequent to the deficient ingestion of water, especially in a hot, dry climate (deserts), or to loss of water through diarrhea, polyuria, profuse sweating, etc. In these cases the total plasma volume diminishes because of the loss of salts and water, the concentration of plasma proteins increases, and the hematocrit reveals an increase in the relative erythrocyte volume; this condition is known as hemoconcentration. The increase in plasma protein concentration can be demonstrated by chemical analysis or (more easily and rapidly, and with sufficient accuracy for clinical purposes) by the changes in the refractometric index or the specific gravity.

Rowntree's terminology has been widely adopted to describe the different alterations in the blood volume. *Normovolemia* is the term used for a normal blood volume, *hypovolemia* for a decreased blood volume, and *hypervolemia* for an increased blood volume. Each of these three states can be subdivided into three, according to the proportions of erythrocytes and plasma. When the ratio of cells to plasma is normal, the blood condition will be normocythemic; it will be polycythemic when the proportion of erythrocytes increases and oligocythemic when this proportion decreases. Therefore with regard to blood volume and the plasma-cell ratio, there are nine possible blood states (Fig. 4).

The blood volume is the result of a dynamic equilibrium. Thus, the plasma volume results from the balance of water outgoing from the blood vessels and incoming from the tissue fluids. The main force which sends water out from the capillaries is the blood pressure, and the principal one for retaining or reabsorbing water is the osmotic pressure of the plasma proteins. After blood is lost by hemorrhage, the blood pressure falls and the plasma proteins attract water from the tissues; in this way the blood volume is rapidly restored. Intravenous injection of saline solution increases the capillary pressure and dilutes the plasma proteins. The blood volume is increased for only a very short time, however; within 30 min. it has returned to the initial level, as water and salts pass out into the tissue fluids and are later eliminated by the kidneys. To obtain a lasting increase in the blood volume it is necessary to transfuse whole blood, or plasma, or serum albumin; in this way water from the tissue fluids will be attracted into the blood vessels, the blood volume will be restored, and consequently the blood pressure will be high. Transfusion of blood, plasma, or serum albumin is a therapeutic measure of great value in traumatic and surgical shock and in cases of severe burns, in which the blood volume

[1] The blood is approximately 8 per cent of the body weight, and it can be calculated by multiplying the body weight by 8 and dividing it by 100. Thus a subject weighing 70 kg. has a blood volume of about 5.6 liters. It is useful to remember this in dealing with cases in which bleeding or blood transfusion must be performed.

diminishes, there is a decreasing venous return to the heart, and the blood pressure falls to a low level. It has also made possible severe and prolonged operations which could not otherwise have been tolerated by the patients. In severe hypovolemia appropriate measures should reaches a maximum of 23 per cent for whole blood and 25 per cent for plasma. The increase in plasma is greater than that of whole blood (oligocythemic hypervolemia) between the eighth and ninth months. The blood volume returns to normal in a few weeks after delivery.

Volume normal or normovolemia
Simple normovolemia
Polycythemic normovolemia
Oligocythemic normovolemia.
Volume decreased or hypovolemia
Simple hypovolemia
Polycythemic hypovolemia
Oligocythemic hypovolemia
Volume increased or hypervolemia
Simple hypervolemia
Polycythemic hypervolemia
Oligocythemic hypervolemia
Cells Plasma

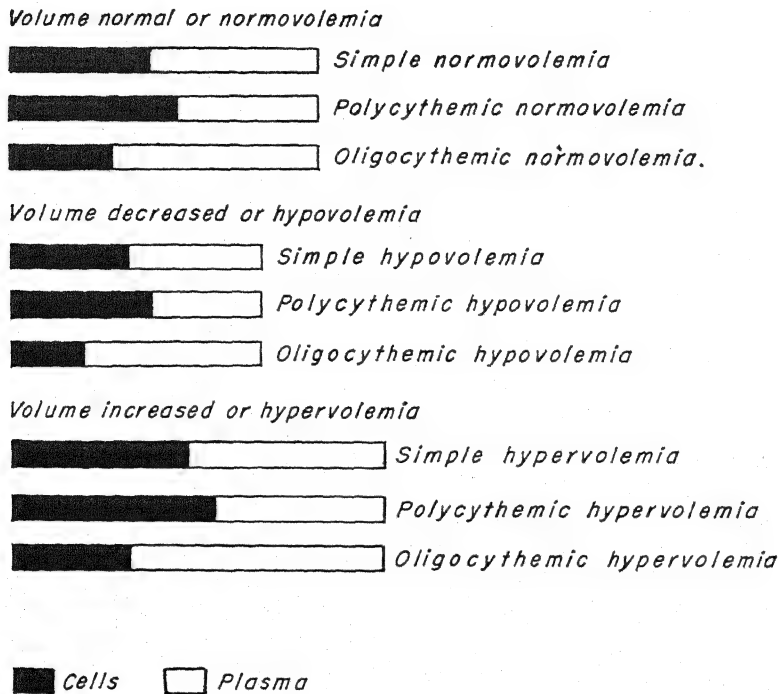



Fig. 4. The nine possible blood states produced by variations in erythrocyte and plasma volumes. (*Rowntree.*) Erythrocyte volume in black; plasma volume in white.

be taken as soon as possible or disturbances in the circulation will become irreversible, and in spite of treatment, the patient will die (see Chap. 7, Hemorrhage and Transfusion).

Physiologic Variations

Changes in the erythrocyte volume can occur rapidly or slowly. Rapid changes are caused by the retention of erythrocytes in the blood stores, especially in the spleen, or inversely, by the release of stored erythrocytes by contraction of the spleen, as occurs in muscular exercise, emotional states, anoxia (lack of oxygen), etc. Slow variations are due to an increase in the formation or destruction of red cells.

Apart from the increase in basal blood volume due to growth, which has already been mentioned, there is an increase caused by pregnancy. This increase has a gradual onset and

After meals or the ingestion of fluid a very small and transitory increase in blood volume can be observed. Water is rapidly taken up by the tissues and then eliminated by the kidneys, so there are only insignificant changes in blood volume even after copious drinking.

An increase in the temperature of the environment produces an increase in the total erythrocyte volume, due to contraction of the spleen, and an increase in total plasma volume, due to reabsorption of water from the tissue fluids. Cold, on the contrary, diminishes the plasma volume, because water passes out into the tissues, especially the skin, muscles, and liver, without there being any change in the total body water. The central nervous system plays a part in regulating this water displacement, but local mechanisms can also be demonstrated by cooling a limited part of the body.

On changing from the prone to the erect position, plasma volume can diminish by as much as 15 per cent in 30 min. This is due to the passage of fluid out into the tissues of the lower limbs. Exercise produces first an increase in the total blood volume and in the red cell volume, due to contraction of the spleen. Afterward water diffuses from the blood plasma into the tissues because of vasodilatation and the increase in the osmotic pressure of the tissue fluids; this latter is produced by the breaking up of large molecules into smaller ones in the course of muscular contraction. Finally sweating provoked by exercise causes loss of water and salts, so the plasma volume diminishes even more and hemoconcentration is more marked.

Altitude, such as is experienced in aviation or climbing a high mountain, produces an increase in the erythrocyte volume due to contraction of the spleen. After a time at high altitude the production of erythrocytes is stimulated and contributes to maintain the polycythemic condition. These changes are caused by the low partial pressure of oxygen.

Pathologic Variations

There are many abnormal conditions in which the blood volume diminishes. Hemorrhage causes a loss of whole blood. There is at first a normocythemic hypovolemia, but water is very rapidly reabsorbed from the tissues, the blood is diluted, and the plasma volume increases. This dilution can be demonstrated by the hematocrit, which shows a relative decrease of the erythrocytes, and by the decrease in the concentration of plasma proteins. Dilution is greater and lasts longer in proportion to the amount of blood lost. Hemorrhage provokes contraction of the spleen, which discharges stored erythrocytes into the circulation; later erythropoiesis (formation of erythrocytes) is stimulated and the normal erythrocyte concentration is restored.

There are three principal mechanisms that cause a decrease in the fluid part of the blood: (*a*) the passage of water and salts into the tissues; (*b*) the loss of water and salts from the body; and (*c*) the loss of plasma.

In traumatic and surgical shock, acute adrenal insufficiency, and diabetic coma, water and salts diffuse from the blood vessels into the tissues; there is a progressive decrease in plasma volume and hemoconcentration (increase in the concentration of erythrocytes and plasma proteins) that can be fatal. Anhydremia and hemoconcentration (polycythemic hypovolemia) are also caused by intense dehydration due to insufficient intake or excessive loss of water, such as occurs in diarrhea, polyuria, profuse sweating, repeated vomiting, high intestinal obstruction, etc. Hemoconcentration is in some cases so marked that it is difficult to extract blood from the depleted veins. The thick, viscous aspect of the blood in the extreme dehydration of cholera is described in all the classic treatises. Polycythemic hypovolemia is present in cases of severe burning, mainly as a result of the loss of plasma (plasmorrhea) into the damaged tissues. Plasmorrhea and hemoconcentration are also observed in pulmonary edema and in dysentery.

In myxedema caused by thyroid insufficiency, both the plasma and erythrocytes are diminished, but there is a certain degree of oligocythemia. In anemia the whole blood volume is only slightly diminished; the decrease in erythrocyte volume is compensated in part by a moderate increase in plasma volume (oligocythemic hypovolemia). The blood volume may be normal in chronic nephritis, but the erythrocyte volume is decreased (oligocythemia). In obesity the blood volume per kilogram is reduced, but when calculated with regard to the body surface it is normal.

Polycythemia vera consists in an increase in whole-blood volume; the erythrocyte volume increases more than the plasma volume, so there is a polycythemic hypervolemia. Chronic anoxia (lack of oxygen) of pulmonary or cardiac origin also causes polycythemia. In acute congestive cardiac insufficiency there is a 20 to 50 per cent increase in plasma volume. Cirrhosis of the liver is also accompanied by an increase in plasma volume. In leukemia the leukocyte and plasma volumes are increased.

Variations in blood volume caused by disease are of vital importance in some cases. Hypovolemia and hemoconcentration, if sufficiently marked and prolonged, cause a decrease in the blood flow through the tissues and consequently an insufficient supply of oxygen (anoxia); anoxia increases the permeability of the capillaries, which accentuates the loss of water and plasma proteins from the blood. Thus a vicious circle is established: plasmorrhea → hypovolemia → hypotension → anoxia → plasmorrhea. If adequate treatment by transfusion of blood, plasma, or serum albumin is not given before the disturbance has progressed too far, death will occur.

An increase in the circulating blood volume increases the work of the heart. An increase in the erythrocyte concentration (polycythemia) increases blood viscosity; hence the peripheral resistance and

the work of the heart are increased, and the circulation tends to become sluggish.

BIBLIOGRAPHY

Blood Volume

BERLIN, N. I., *et al.*, *New England J. Med.*, **247**, 675, 1952.

GIBSON, J. G., *Ann. Int. Med.*, **14**, 11 and 2014, 1941.

GIBSON, J. G., and W. A. EVANS, *J. Clin. Investigation*, **16**, 301 and 317, 1937.

GREGERSEN, M. J., *et al.*, *J. Lab. & Clin. Med.*, **23**, 423, 1938; **29**, 1266, 1944.

GREGERSEN, M. J., *et al.*, *Am J. Physiol.*, **113**, 54, 1935; **120**, 494, 1937; **121**, 284, 1938; **125**, 142, 1939.

GREGERSEN, M. J., and J. L. NICKERSON, *J. Applied Physiol.*, **3**, 328, 1950.

MENEELY, J. R., E. B. WELLS, and P. F. HAHN, *Am. J. Physiol.*, **148**, 531, 1947.

ROWNTREE, L. G., and G. E. BROWN, "The Volume of the Blood and Plasma in Health and Disease," Saunders, Philadelphia, 1929.

CHAPTER 3

The Erythrocytes[1]

Functions. The red blood cells or erythrocytes, which give the blood its typical scarlet color, have the following functions:

1. They carry oxygen and carbon dioxide.
2. They take part in the regulation of the acid-base equilibrium.
3. Their coloring matter, hemoglobin, gives rise to the bile pigments and allied substances.

Shape and size. Under the microscope the erythrocytes are seen to be biconcave disks, reddish yellow in color, flexible, and elastic; their shape can be passively changed by outside forces, but afterward they recover their original aspect. The erythrocytes of fishes, amphibians, reptiles, and birds have a nucleus and usually are biconvex with an elliptic contour. The erythrocytes of mammals have no nucleus and are discoid in form, except in some Camelidae (llama, guanaco) in which they have an elliptic shape. In man the erythrocytes are discoid except for a few exceptional cells which have an oval or elliptic shape (ovalocytosis). In certain subjects (Africans and Mediterraneans) sickle-shaped erythrocytes are found (drepanocytosis).

The mammalian erythrocyte is well adapted to its respiratory function. The lack of a nucleus, the small size, and the discoid biconcave shape allow the transport of a relatively large amount of hemoglobin, distributed near the surface of the cell so as to facilitate gaseous exchange. These nonnucleated erythrocytes are highly differentiated cells that have lost their reproductive power and consume much less oxygen than the nucleated erythrocytes. The biconcave form is a great advantage (Ponder), as for a given mass it assures a maximum surface and the proximity to this surface of even the most deeply placed molecules.

The surface of each erythrocyte is 128 sq. μ (120 to 140 sq. μ). The total surface of all the erythrocytes in the organism has been estimated at 3,500 sq. m., nearly 2,000 times the body surface. Bürker has calculated that the amount of hemoglobin per square micron (sq. μ) of erythrocyte surface is almost constant (about 32×10^{-14}) in all mammals, in spite of differences in size and concentration of erythrocytes and in hemoglobin concentration.

The size of the erythrocytes can be determined by measuring the diameter, the thickness, or the volume of the cell. The average diameter in dry preparations varies with the technique used for preparing the specimens; it is 7.2 μ (6 to 8 μ) according to Price-Jones, or 7.5 μ according to Wintrobe; in plasma it is slightly larger, approximately 8.5 μ. CO_2 and acidosis increase the diameter, because water and Cl enter into and swell the erythrocyte. Therefore, in venous blood the erythrocytes are larger (0.5 μ) than in arterial blood. Size-frequency curves (Fig. 5), obtained by measuring large numbers of erythrocytes, show that the great majority have a size near the mean (normocytes), but there are a few small cells (microcytes) and a few large cells (macrocytes).[1] In cases of pernicious anemia the average diameter is larger than the normal and the dispersion in size is greater (Figs. 5 and 6), because of the presence of large erythrocytes developing from primitive cells called megaloblasts. In some types of

[1] The erythrocytes were accurately described for the first time by Leeuwenhoek (1673). Hemoglobin was obtained in crystallized form by Funke (1851), and its respiratory function was demonstrated by Hoppe Seyler (1867). Vierordt made the first red cell counts in 1852. Neumann (1868) observed the formation of erythrocytes in the bone marrow.

[1] Price-Jones, C., *J. Path. & Bact.*, **40**, 503, 1935.

anemia there is a great disparity in the size of the erythrocytes (anisocytosis), and sometimes they have abnormal and varied shapes (poikilocytosis).

The normal thickness of the red cells varies from 1.8 to 2.2 μ, but in pathologic cases

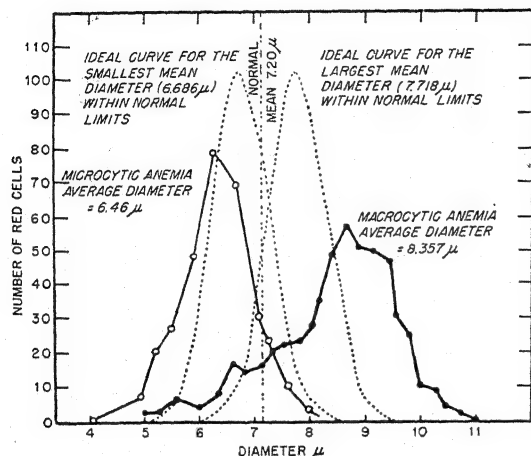



Fig. 5. Distribution curve of erythrocyte diameter in several blood conditions. (*Whitby, L. E. H., and C. J. C. Britton, "Disorders of the Blood," J. & A. Churchill, London, 1939.*)

(hemolytic anemia, congenital hemolytic jaundice) it can increase considerably and the erythrocytes take on a spheroid shape (spherocytes). This shape is observed *in vitro* when the erythrocytes are placed in a hypotonic solution. In these cases the erythrocytes increase in thickness but not in diameter, as can be seen by dividing the diameter by the thickness; the average quotient in normal subjects is 3.4 (2.5 to 4) (Fig. 6). The easiest way of measuring the true size of the erythrocytes is to determine the mean corpuscular volume (see below).

Concentration. (Fig. 7.) Well-nourished adult men have about 5,400,000 erythrocytes per cu. mm. (4,500,000 to 6,000,000); the corresponding figure for women is 4,800,000 (4,000,000 to 5,500,000) per cu. mm. The newborn have a high concentration of large erythrocytes; in the course of the first few weeks of life the concentration diminishes and normal-sized erythrocytes appear. During childhood the concentration of erythrocytes remains low, and rises during puberty to the adult level. During pregnancy there is a slight decrease in the concentration due to an increase in plasma volume; the drop is usually not greater than 15 per cent of the normal figure.

Variations in the erythrocyte concentration are produced by one of the following mechanisms:

1. Excessive or insufficient production.
2. Excessive destruction.
3. Loss of erythrocytes (hemorrhage).
4. Increase or decrease in plasma volume.
5. Rapid mobilization of erythrocytes stored in the spleen.

An increase in erythrocyte concentration is called "polycythemia," and a decrease "oligocythemia." Three types of polycythemia can occur:

1. Relative polycythemia, when the increase in concentration is due to loss of plasma or water (hemoconcentration).
2. Transitory polycythemia, when erythrocytes are rapidly discharged from the blood reservoirs.
3. Absolute polycythemia, when the total amount of erythrocytes in the blood is increased.

The influence of oxygen and nutritional factors which regulate erythrocyte concentration will be discussed in Chap. 5.

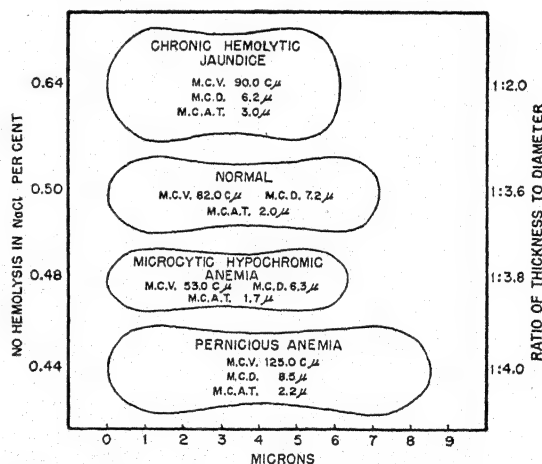



Fig. 6. Volume, diameter, and thickness in relation to fragility of erythrocytes. *M.C.V.*, mean corpuscular volume; *M.C.D.*, mean corpuscular diameter; *M.C.A.T.*, mean corpuscular average thickness. (*Whitby, L. E. H., and C. J. C. Britton, "Disorders of the Blood," J. & A. Churchill, London, 1939.*)

Chemical composition. (See Tables 1 and 2, pages 8 and 9.) Water makes up about two-thirds of the mass of the erythrocyte. Hemoglobin constitutes 90 to 95 per cent of the solid content, about

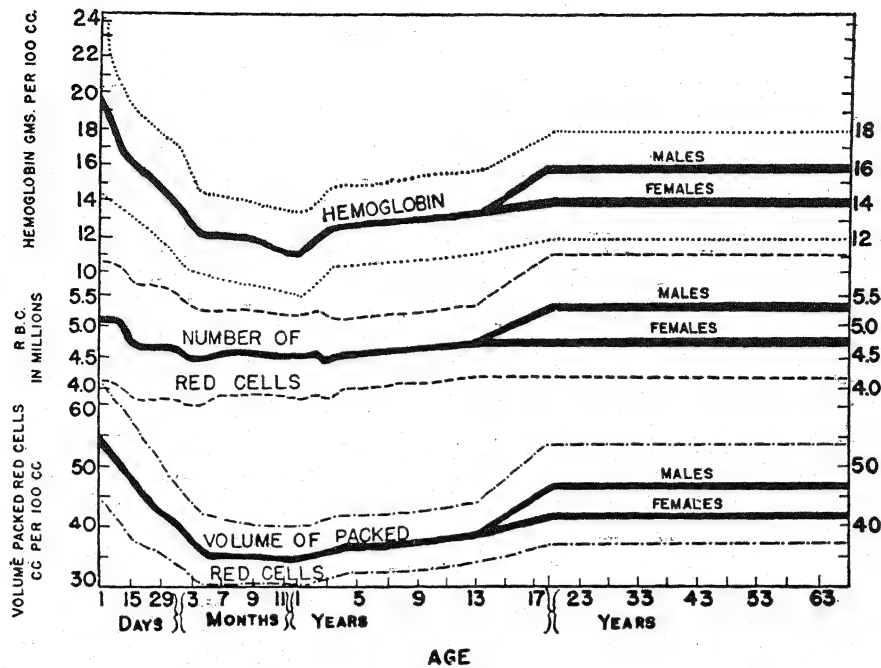



FIG. 7. Normal hemoglobin and erythrocyte concentration and volume of packed red cells at different ages. The range of variation is given by the dotted or broken lines. (*Wintrobe, M. M., "Clinical Hematology," Lea & Febiger, Philadelphia, 1942.*)

Table 3. Normal Values for Human Erythrocytes at Different Ages

Age	*Erythrocyte concentration, millions per cu. mm.*	*Hemoglobin, gm. per 100 cc.*	*Volume of erythrocytes centrifuged from 100 cc. of blood, cc.*	*Erythrocyte standards*		
				Mean corpuscular volume, cu. μ	*Mean corpuscular hemoglobin, γγ*	*Mean corpuscular hemoglobin concentration, %*
1 day	5.3	23.0	51.0	106	38	36
4 to 8 days	4.7	19.0	47.0	103	36	35
3 to 5 months	4.5	12.2	36.0	80	27	34
1 year	4.5	11.2	35.0	78	25	32
5 years	4.6	12.6	37.0	80	27	34
11 to 15 years	4.8	13.4	39.0	82	28	34
Adults						
Women	4.8	13.8 ± 2	42.0	87 ± 5	29 ± 2	34 ± 2
Men	5.4	15.4 ± 2	47.0	87 ± 5	29 ± 2	34 ± 2

Modified from Wintrobe.

34 gm. in 100 cc. of the circulating erythrocytes. Other proteins make up from 0.5 to 1 per cent of the cell mass. Nucleoproteins have been found in the stroma. Phospholipids are in higher concentration than in plasma; there are also cholesterol esters and free cholesterol, cerebrosides, and small quantities of neutral fats. Reducing substances which are not glucose or any other sugar are found in the erythrocytes. In the erythrocytes of man and several animals, K is the predominant cation; in other species (cat, dog) Na predominates. The erythrocyte has a higher nonprotein N concentration than plasma. The erythrocytes have catalase, which can be demonstrated by adding red cells to hydrogen peroxide, whereupon bubbles of molecular

oxygen are set free. Hemoglobin acts as a peroxidase and transports oxygen from hydrogen peroxide to a substrate. This reaction is used to reveal the hidden presence of blood in feces and in body fluids and secretions in pathologic cases. Substrates which take on a typical color on oxidation are employed (Mayer's reagent, tincture of guaiac, benzidine).

Table 4. Erythrocyte Volume in Different Types of Anemia

Type of anemia	*Mean corpuscular volume, cu. μ*	*Mean corpuscular hemoglobin, γγ*	*Mean corpuscular concentration, %*	*Mean diameter of erythrocytes, μ*
Macrocytic	95–160	30–52	31–38	7.8–8.5
Normocytic	80–94	27–32	33–38	6.8–8.0
Microcytic	70–79	22–26	31–38	6.5–8.5
Microcytic hypochromic	50–71	14–21	21–29	5.8–6.5

Absolute corpuscular values. Three measurements are necessary to determine the condition of the erythrocytes in health and in disease. These measurements are (*a*) the erythrocyte volume (with the hematocrit); (*b*) the red cell count; (*c*) the hemoglobin concentration. From the data obtained by these determinations, the following values can then be calculated:

1. Mean corpuscular volume (MCV), *i.e.*, the average volume of each erythrocyte expressed in cubic microns (cu. μ):

$$\frac{\text{Volume of erythrocytes (in cc.) in 1,000 cc. of blood (hematocrit)}}{\text{Erythrocytes per cu. mm. (in millions)}} = 87 \text{ (80 to 94) cu. } \mu$$

2. Mean corpuscular hemoglobin (MCHb), *i.e.*, average amount of hemoglobin in each erythrocyte in micro-micrograms ($\gamma\gamma$):

$$\frac{\text{Hemoglobin (in gm.) in 1,000 cc. of blood}}{\text{Erythrocytes per cu. mm. (in millions)}} = 29 \text{ (27 to 32) } \gamma\gamma$$

3. Mean corpuscular hemoglobin concentration per cent (MCHbC), *i.e.*, the amount of hemoglobin in 100 cc. of erythrocytes:

$$\frac{\text{Hemoglobin (in gm.) in 100 cc. of blood}}{\text{Volume of erythrocytes in 100 cc. of blood (hematocrit)}} \times 100 = 34 \text{ (33 to 38) gm.}$$

Anemias are classified according to the size of the erythrocytes into macrocytic, normocytic, and microcytic. When the corpuscular hemoglobin is also diminished, the anemia is called hypochromic.

The mean corpuscular thickness is calculated from the mean corpuscular volume (*MCV*) and the mean corpuscular diameter (*MCD*) in the following way:

$$\frac{MCV}{\pi(MCD/2)^2} = 2.1 \text{ (1.7 to 2.5) } \mu$$

Hematic indexes: These are values given by comparing the result obtained in a particular case with the normal value. These indexes should be abandoned in favor of the absolute values.

The color index is calculated in the following way:

$$\frac{\text{Hemoglobin, \% of normal (gm. in 100 cc.} \times 6.9)}{\text{Erythrocytes, \% of normal (millions per cu. mm.} \times 20)} = 1 \text{(0.9 to 1.1)}$$

The normal erythrocyte count is considered as 5,000,000 and the normal hemoglobin 14.5 gm. per cent (Wintrobe).[1]

The color index is high in pernicious anemia, which has for this reason been wrongly called "hyperchromic" anemia. The erythrocyte contains more hemoglobin because it is larger, but the hemoglobin concentration is normal (34 gm. per 100 cc. of cells). Macrocytic and microcytic anemias are hypochromic when the corpuscular hemoglobin and the hemoglobin concentration are diminished.

The saturation index is calculated thus:

$$\frac{\text{Hemoglobin, \% of normal (gm. in 100 cc.} \times 6.9)}{\text{Volume of packed cells (of 100 cc.)} \times 2.3} = 1 \text{(0.8 to 1.2)}$$

The volume index is calculated thus:

$$\frac{\text{Volume of packed cells (of 100 cc.)} \times 2.3}{\text{Erythrocytes per cu. mm. (in millions)} \times 20} = 1 \text{(0.9 to 1.1)}$$

[1] The average normal figure in Buenos Aires is 5,400,000 and 15.6 gm. per cent.

Agglutination. The erythrocytes of shed blood form groups, similar to piles of coins, known as "*rouleaux.*" *Rouleaux* formation is seen within the vessels only when the blood circulates very slowly. If the shed blood is shaken up, the erythrocytes separate from each other. In some cases, as in pregnancy, acute infections, malaria, etc., the formation of *rouleaux* is marked and the piled erythrocytes are separated from the plasma, but on shaking, the *rouleaux* break up. These are phenomena of pseudoagglutination.

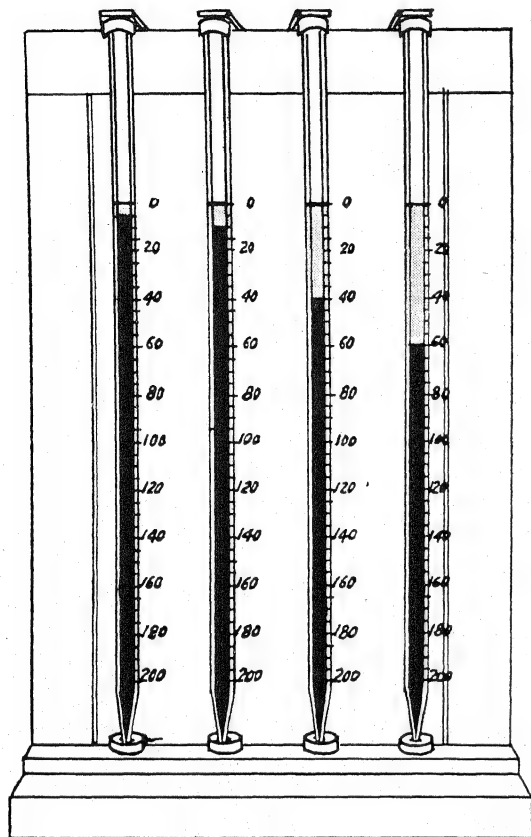



Fig. 8. Westergren's technique for determining sedimentation rate. Normal in the two tubes on the left and accelerated in the two on the right.

When blood from one subject is mixed with serum from another subject, the erythrocytes of the first subject may remain dispersed as in their own plasma, or they may aggregate in clumps which are not broken up by shaking; this phenomenon is called agglutination. The serum in this case has an agglutinating factor, or agglutinin, for the erythrocytes, known as hemagglutinin; more specifically it is called "isohemagglutinin," because it acts on the erythrocytes of subjects belonging to the same animal species. The presence or absence of agglutinin in serum and of the corresponding factor (agglutinogen) in the erythrocytes has led to the classification of human beings in four principal blood groups (see "Blood groups," Chap. 7). The serum of an animal species can also have agglutinins for the erythrocytes of other species; those are called heteroagglutinins. The repeated injection of erythrocytes of one species into an animal of another species provokes the formation of heteroagglutinins in the blood of the latter; as this is a process of immunity, the agglutinins so formed are known as immune hemagglutinins.

The erythrocytes are agglutinated by hemolytic serum obtained by immunization (see "Hemolysis"); by some snake venoms, *e.g.*, those of *Bothrops;* by several animal and plant toxins; and by some chemical agents.

Suspension stability. Erythrosedimentation. When blood in which clotting is prevented is left standing in a long tube placed vertically, the erythrocytes sink to the bottom, because their specific gravity is greater than that of plasma. Above the red cells a thin layer of leukocytes is deposited, covered by plasma in which the platelets are suspended. Blood is a suspension of cells in plasma, the stability of which can be measured by the velocity of sedimentation of the erythrocytes. There are several methods for determining the suspension stability of the blood; Westergren's technique is the most simple and commonly used.

Blood is collected, mixed with one-fifth its volume of a 3.8 per cent sodium citrate solution, and placed in a graduated pipet 300 mm. long. The pipet is left standing, and at the end of 1 hr. the length of the column of separated plasma is read (Fig. 8). Wintrobe's hematocrit can be used for this purpose, leaving the tube and anticoagulant standing for 1 hr. before centrifugation. Heparin (1 mg. per cc. of blood) is also used instead of citrate as an anticoagulant.

The sedimentation of corpuscles suspended in a fluid depends on the specific gravity and mass of the corpuscle and on the resistance that the fluid exerts on the surface of the corpuscle. The interaction of the factors that play a part is expressed in Stokes's law:

$$V_s = \frac{2r^2(S - S_1)g}{9\eta}$$

where V_s is the sedimentation velocity, r is the radius of the corpuscle, S is the specific gravity of the corpuscle, S_1 is the specific gravity of the fluid, g is the gravitation constant, and η is the viscosity coefficient of the fluid.

For sedimentation to take place $S - S_1$ must have a positive value, as occurs in blood, where the corpuscles are heavier than the plasma.

The sedimentation rate is greater when the corpuscles are larger; the aggregation of erythrocytes into *rouleaux* accelerates their sedimentation. Variations in the sedimentation rate depend almost exclusively on plasmatic factors, and little or not at all on the erythrocytes. Red cells from normal blood sink rapidly when placed in plasma from blood with a high sedimentation rate, and red cells from blood with a high sedimentation rate sink at a normal rate in normal plasma. The most important factor is the concentration of fibrinogen, to which the sedimentation rate is directly proportional; globulin concentration is not so important, and serum albumin even less. When fibrinogen concentration increases, the erythrocytes conglomerate into larger *rouleaux* and thus sink more rapidly. Apparently the negative charge, which keeps the erythrocytes dispersed, is diminished in this case; also there seems to be an increase in the surface tension and hydrophilia of the erythrocytes. When fibrinogen concentration diminishes, as in defibrinated blood and in some cases of shock, the erythrocyte sedimentation rate is very low. Sedimentation rate increases when the erythrocyte concentration diminishes, *e.g.*, in anemia.

The blood plasma of children has a low globulin concentration and a low sedimentation rate; the newborn child has a rate of 0.2 to 0.5 mm. per hr. With age the rate increases to an average of 3.3 (2 to 7) mm. per hr. in adult men. The plasma of adult women has a higher globulin concentration and therefore a faster sedimentation rate; the average is 7.5 (3 to 10) mm. per hr. The sedimentation rate increases during pregnancy up to 45 mm. toward the end, and diminishes gradually after delivery to 41 mm. in 1 month and 20 mm. in 2 months (Fåhraeus, 1918). Oxygen accelerates erythrosedimentation because it increases the size of the erythrocytes; CO_2 retards sedimentation.

In cyanosis (lack of oxygen in the blood) the suspension stability of the blood is greater than normal. In polycythemic conditions erythrosedimentation is slow. In acute general infections (pneumonia, rheumatic fever, etc.), local infections (abscesses, arthritis, pelviperitonitis, etc.), the acute periods of chronic infections (tuberculosis, leprosy, etc.), and malignant tumors, there is an increased sedimentation rate. The changes in sedimentation rate are not specific for any particular disease, but nevertheless they are very useful in medical practice; because an increase reveals the existence of inflammation or tissue destruction, it can therefore be used to follow the course of such a process, and to establish its diagnosis and prognosis.

HEMOLYSIS

Description. Normally the hemoglobin of the erythrocytes does not diffuse out into the plasma or into isotonic saline solutions. Several physical and chemical agents can act on the erythrocytes so as to set free the hemoglobin, which then diffuses into the fluid surrounding the red cells; the latter seem to disappear and the fluid becomes darker and transparent. This dissolution of the erythrocytes is called "hemolysis," and the blood is said to be "hemolyzed" or "laked." By centrifugation of laked blood, especially if certain salts are added, a small amount of cell residue is obtained.[1] This is the stroma of the erythrocytes, consisting in the protein and lipid framework which holds the hemoglobin and electrolytes.[2]

Permeability of the erythrocyte. The erythrocyte membrane is easily permeable to the anions Cl^-, CO_3H^-, and to the cation H^+; also to glucose, amino acids, urea, and NH_4. It is not permeable to proteins and does not let the plasma proteins enter the erythrocyte, nor the hemoglobin diffuse into the plasma. It was thought to be impermeable to Na^+, K^+, Ca^{++} and Mg^{++}, but by using radioactive K and by other methods it has been possible to demonstrate that small amounts of these ions pass through the membrane.

[1] The erythrocytes are separated by centrifugation and washed twice in saline solution. To the mass of cells obtained by centrifugation, two volumes of distilled water saturated with ether are added. Then, drop by drop, a 1 per cent solution of potassium acid sulfate is added until the solution has the same opacity as blood. By centrifugation the stroma of the erythrocytes is separated and can be washed with distilled water.

[2] Hemoglobin is retained by the erythrocyte even after it has been damaged mechanically or divided into fragments, if the cells are in an isotonic solution.

The fragility (resistance) of erythrocytes. The properties of the erythrocyte are sometimes explained by considering it as a minute osmometer surrounded by a semipermeable membrane, made up of a mosaic or of concentric layers of proteins, lipids, and cholesterol. The membrane of the erythrocyte is not truly semipermeable, as it lets through not only water but also some substances in solution; it has a selective permeability. The contents of the erythrocyte have an osmotic pressure equivalent to that of blood plasma. When the plasma is diluted, water penetrates into the erythrocytes, which swell and take on a spheroid shape; finally hemoglobin diffuses out as if the membrane had been broken or had become permeable.

The erythrocytes retain their size and shape when suspended in an isotonic solution. Such a solution can be prepared by dissolving NaCl in water—0.95 per cent for men, and 0.93 per cent for women.[1] If erythrocytes are suspended in NaCl solutions decreasing by 0.02 per cent from 0.6 per cent concentration, hemoglobin will begin to diffuse out of the cells and color the fluid at a concentration of 0.48 to 0.40 per cent; the more fragile erythrocytes cannot resist this degree of hypotonicity (minimum resistance). Total hemolysis will be observed at concentrations of 0.34 to 0.30 per cent, when even the most resistant cells let their hemoglobin escape (maximum resistance). The spheroid erythrocytes are fragile, while flat, elongated erythrocytes are resistant. The resistance of the cells is modified by the temperature, pH, and ionic equilibrium of the solution. Special techniques have been devised to measure the fragility or resistance of the erythrocytes taking all these factors into account.[2] A more accurate study requires the determination of the amount of hemoglobin diffusing out at each concentration level.

The fragility of the erythrocytes is considerably increased in congenital hemolytic jaundice (Fig. 6) and in some types of purpura; frequently the fragile cells have a spheroid shape. The resistance of the erythrocytes is increased in pernicious anemia, and markedly increased in infantile erythroblastic anemia; it is also increased in jaundice due to obstruction of the bile ducts. Red cells are not hemolyzed in an isotonic solution of sucrose. On the other hand, a solution of urea hemolyzes the erythrocytes because urea diffuses into the cells so that its concentration is the same within and without the erythrocyte; therefore the erythrocytes swell out and hemolysis occurs.

Physical and chemical hemolytic agents. Hemolysis can be produced by intense and prolonged shaking; electrical discharges; repeated freezing and thawing; heat (65°C.); organic solvents, such as ether, alcohol, toluene, chloroform, benzene, etc.; alkalis, especially NH_3; bile salts; saponin, which is very active but can be neutralized by cholesterol; toluylene-diamine; phenylhydrazine; and several arsenic compounds. Some of these substances, *e.g.*, phenylhydrazine, have been used in the treatment of polycythemia to diminish the excessive concentration of erythrocytes.

Many plant and animal substances and toxins have hemolytic activity. Among these are some of the snake venoms, which contain phosphatidases; they attack lecithin and cephalin, and on separating the oleic acid from the molecule, powerful hemolytic substances are formed, lysolecithin and lysocephalin (Delezenne and Fourneau, Levene). Cobra (*Naja tribudians*) venom acts directly on the phosphatides of the erythrocytes. Other snake venoms act on the lecithin in the plasma, which must be added when experimenting with cells suspended in saline solution. Some of the bacteria, *e.g.*, streptococci, *Bacillus perfringens*, have hemolytic activity.

Hemolytic serum. (See Chap. 9.) Blood serum hemolyzes the erythrocytes of other species, and can acquire or increase this activity by immunization. For example, repeated injections of sheep erythrocytes into a rabbit produce strong hemolytic activity in the serum of the rabbit against sheep erythrocytes. The erythrocytes act as an antigen and provoke the active formation of an antibody (hemolysin). This hemolysin is made up of two parts, the amboceptor (Ehrlich), or sensibilisatrix (Bordet), which is specific for the erythrocytes used as antigen, and the complement (Ehrlich), or alexin (Bordet), which is not specific. By heating the hemolytic serum at 56°C. for 30 min., the complement is destroyed, while the amboceptor resists this temperature. Hemolytic activity can be restored by adding fresh serum of any species, *e.g.*, guinea pig, and thus replacing the destroyed complement.

[1] Care must be taken to avoid loss of CO_2; otherwise the isotonic concentration of NaCl is 0.9 per cent.

[2] Hamburger recommends the use of Na_2SO_4 instead of NaCl.

Plasma, or serum, usually does not hemolyze the erythrocytes of the same species, but in some cases there are isohemolysins, which can be responsible for accidents in blood transfusion.

In paroxysmal hemoglobinuria, hemolytic crises occur when the patient is exposed to cold, or exercises, or suffers an intense emotional stress. If blood is drawn and kept warm, hemolysis does not occur, but if it is cooled to 5°C. and then warmed to 38°C., there is hemolysis (Donath). Isohemolysins have been found in the blood of patients with hemolytic jaundice.

BIBLIOGRAPHY

Erythrocytes

GRANICK, S., The Chemistry and Functioning of the Mammalian Erythrocyte, *Blood*, **4**, 404, 1949.

HADEN, R. L., *Internat. Clin.*, **1**, 68, 1935.

KNISELY, M. H., *et al.*, Sludged Blood, *Science*, **106**, 438, 1947.

PARODI, A. S., *Rev. Soc. argent. de biol.*, **6**, 426, 1930 (hemoglobin and erythrocytes).

PONDER, E., "Hemolysis and Related Phenomena," J. & A. Churchill, London, 1948.

WINTROBE, M. M., The Erythrocytes in Man, *Medicine*, **9**, 195, 1930.

WINTROBE, M. M., "Clinical Hematology," 3d ed., Lea & Febiger, Philadelphia, 1951.

Erythrosedimentation

FÅHRAEUS, R., *Physiol. Rev.*, **9**, 241, 1929.

HAM, T. H., and F. C. CURTIS, *Medicine*, **17**, 447, 1938.

WESTERGREN, A., *Klin. Wchnschr.*, **2**, 1359, 1922.

Anemia and Polycythemia

HAM, H., W. B. CASTLE, *et al.*, *New England J. Med.*, **243**, 860, 1950.

CHAPTER 4

Hemoglobin and Derived Pigments

HEMOGLOBIN

Functions. Hemoglobin, the respiratory pigment of the blood, has the following functions: (*a*) it takes up oxygen in the lung and transports it in the circulating blood, giving it up to the tissues; (*b*) it helps to carry CO_2 from the tissues to the lung; (*c*) it takes part in the maintenance of the acid-base equilibrium; (*d*) it is the source of bilirubin, which gives rise to urobilin. The color of the blood is due to hemoglobin and to its combinations and derived pigments.

Localization. Hemoglobin is found in several animal tissues. In the blood of vertebrates it is found exclusively in the erythrocytes, of which it is a constant and typical constituent. This localization within the red blood cells has many advantages (Barcroft); thus blood can transport sixty times the oxygen it could carry if it were made up only of plasma. Nevertheless, it has been possible to replace, for a short time, up to 60 per cent of the blood of dogs by a 6 per cent solution of hemoglobin.[1] Hemoglobin that escapes from the cells into the plasma either is picked up by the cells of the reticuloendothelial system, where it is metabolized, or else is eliminated by the kidney.

Structural relations with other pigments. The most widely distributed pyrrolic respiratory pigments are the "hemes," one of which is hemoglobin. Its fundamental chemical nucleus is a pyrrole group; four of these groups unite to form porphyrins. Porphyrins combine with metals Fe, Cu, Co, etc., and form metalloporphyrins. Hemes are ferroporphyrins, *i.e.*, compounds of porphyrin and iron. One of the porphyrins, protoporphyrin 9, type III, by combining with iron forms heme, which joined to a globin gives hemoglobin. A magnesium porphyrin of the isomeric type III is the fundamental group of chlorophyll, the green pigment of plants.

[1] Amberson, W. R., *et al.*, *Am. J. Physiol.*, **116**, 1, 1936; **117**, 230, 1936.

Hemoglobin is a compound of heme and protein (globin). In red-muscle fibers and in cardiac muscle there is myohemoglobin similar to, but not the same as, hemoglobin. In many, probably in all, animal and plant cells there are compounds of heme, called cytochromes (Keilin). These exist in reduced and oxidized forms, as they can take up and lose electrons, thus taking part in oxidation-reduction reactions. Reduced cytochrome takes up oxygen and releases it to the cells. The amount of cytochrome in a tissue is proportional to the respiratory activity of the cells.

By means of spectroscopic studies it has been possible to identify four different cytochromes, known as α, α_3, β, and c.[1]

The cytochromes form part of the oxidation-reduction mechanisms of the tissues. They do not take up oxygen directly but by means of cytochrome-oxidase (indophenoloxidase). Cyanides and other toxic substances (SH, CO, anesthetics) inhibit cytochromeoxidase and suppress its action in tissue oxidation. Cytochrome is oxidated not directly but through the activity of tissue dehydrogenases, which activate hydrogen so that it combines with cytochrome, the latter acting as a hydrogen acceptor. In this way cytochrome behaves as an intermediary in the transfer

[1] Keilin, D., and W. Hartree, *Proc. Roy. Soc., London*, *s.B.*, **127**, 167, 1939.

of oxygen from hemoglobin and plasma to the oxidizable substrate of the cells.

Heme has also been found in certain enzymes, such as peroxidase and catalase.

Physical properties. Hemoglobin (especially in its oxygenated form, oxyhemoglobin) crystallizes, more or less easily according to the animal species from which it comes, into rhombic prisms or needles of the rhombic system. In some species the crystals are typical tetrahedrons (guinea pigs) or plates (squirrels) (Fig. 9). Hemoglobin in water forms colloidal dispersions which do not dialyze nor ultrafilter. It is a dextrorotatory ampholyte, with the iso-electric point at pH 6.78. The size of its molecule can be measured in several ways: (*a*) by testing whether or not it passes through a membrane, the pores of which have been measured by the velocity of diffusion of water through them; (*b*) by ultracentrifugation at high speed (Svedberg); (*c*) by the osmotic pressure of its solutions (Adair).

The spectroscopic method is of fundamental importance in the study of hemoglobin, its combinations, and the pigments derived from it. Oxyhemoglobin shows two bands in the absorption spectrum, situated between the D and E Frauenhofer lines (Fig. 10). The α band is the narrower, darker, and more clearly marked of the two and is placed at 5760 A; the β band is wider and has its middle point at 5413 A. The spectrum of carboxyhemoglobin (COHb) has two similar bands slightly displaced toward the violet end, the α at 5680 A and the β at 5390 A.

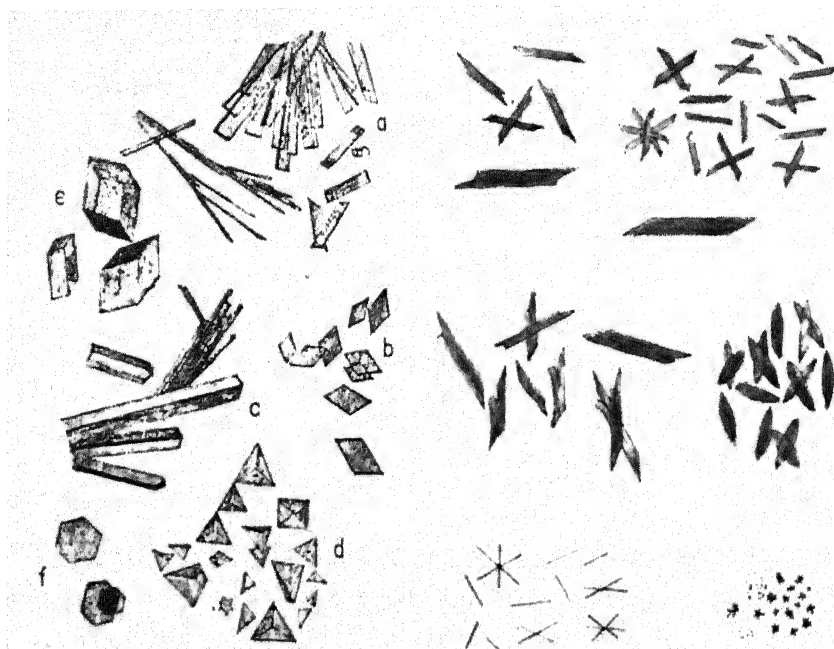



Fig. 9. Left: oxyhemoglobin crystals. *a* and *b*, man; *c*, cat; *d*, guinea pig; *e*, hamster; *f*, squirrel. Right: crystals of hemin or hematin chlorhydrate (Teichmann's crystals).

Oxyhemoglobin loses oxygen by the action of reducing agents, such as sodium hyposulfite ($Na_2S_2O_4$), and becomes reduced hemoglobin; both the bands in the absorption spectrum disappear and are replaced by a single one which is wider, has less distinct margins, and has its center placed at 5560 A. Hemoglobin, its compounds, and the pigments derived from it absorb violet and ultraviolet light (Soret's band).

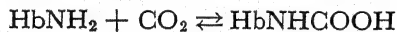
Chemical properties. Hemoglobin is a conjugated protein, made up of 4 per cent heme and 96 per cent globin, which is a histone. Hemoglobin contains C, N, H, O, S, and Fe; the proportion of Fe is fairly constant, about 0.34 per cent. Hemoglobins from different species have the same fundamental structure and properties, but they differ in some of their properties (solubility, readiness to crystallize and shape

of the crystals, velocity of combination with O_2 or CO, amino-acid and S content, etc.). Heme is the same in all hemoglobins, but the globin differs from one species to another. The globin of fetal hemoglobin is not the same as that of adult hemoglobin.

If it is considered that there is at least one atom of Fe in each molecule, the minimum weight possible would be 16,400. The molecular weight of hemoglobin has been determined and found to be about four times this figure—63,000 by ultracentrifugation (Svedberg), and 66,400 by osmotic pressure (Adam). The hemoglobin molecule, therefore, contains four atoms of Fe and is made up of four protein units of 16,400 molecular weight each. This high molecular weight, and the consequently great size of the molecule, are the cause of the colloidal nature of its dispersions. Myohemoglobin has a molecular weight of 16,800 to 17,000; therefore there is only one atom of Fe in each molecule.

Hemoglobin rapidly forms well-defined reversible combinations with oxygen and with carbon monoxide, the substance that is the cause of coal-gas poisoning. These combinations are dependent on the partial pressure of the gas, the temperature, the pH, and the electrolytes in the solution. This property is of primary importance in the respiratory function of hemoglobin, *i.e.*, the capacity to take up, transport, and deliver oxygen; it will be studied in the section on respiration. Each atom of Fe in hemoglobin can combine with one molecule (2 atoms) of oxygen or one molecule of carbon monoxide. (See Chap. 28, The Transportation of Oxygen by the Blood.)

Hemoglobin also combines with CO_2. Globin, not heme, takes part in this combination, which results in a carbamino compound.

$$HbNH_2 + CO_2 \rightleftarrows HbNHCOOH$$

This is one of the ways in which hemoglobin takes part in the transport of CO_2 (see Chap. 30).

Hemoglobin can also combine with nitric oxide (NO), hydrogen sulfide (H_2S), and hydrogen arsenide (H_3As).

Hemoglobin acts as a peroxidase and sets free molecular oxygen from oxygen peroxide. This oxygen can react on tincture of guaiac or on benzidine and gives a colored compound. The reaction is not specific for hemoglobin, but is useful to detect the presence of hemoglobin, and therefore of blood, in feces and other bodily excretions or secretions.

Studies begun by Pauling in 1949 have demonstrated the existence of several types of hemoglobin in man, *e.g.*, HbA (A_1, A_2) found in normal adults; HbF or fetal Hb; HbS found in the sickle-shaped erythrocytes in cases of Cooley's sickle-cell anemia; HbC and HbD.

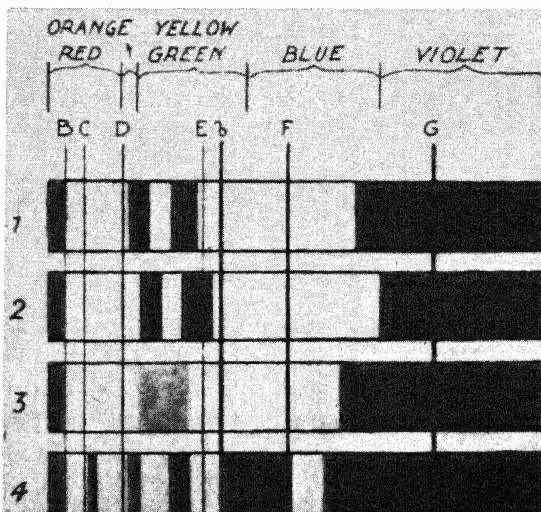



FIG. 10. Absorption spectra of hemoglobin and some of its derivatives. 1, oxyhemoglobin; 2, carboxyhemoglobin; 3, reduced hemoglobin; 4, methemoglobin.

Oxygen capacity. The reversible combination of hemoglobin with oxygen is known as oxygenation. It is not a true oxidation and its Fe remains in the ferrous (Fe^{++}) state, not passing to the ferric (Fe^{+++}) state. It is usually admitted that 1 gm. of hemoglobin can take up 1.34 cc. of oxygen.[1] This is only an approximate figure, because of the difficulty in obtaining pure hemoglobin; moreover, on drying, hemoglobin loses part of its oxygen capacity. The best samples of crystallized hemoglobin so far obtained have about 95 per cent the oxygen capacity of hemoglobin in fresh blood. The oxygen capacity of 100 cc. of blood is on an average 21 cc. in man (15.4 gm. Hb × 1.34 cc. O_2) and 18.5 cc. in woman (13.8 gm. Hb × 1.34 cc. O_2). This maximum amount is taken up in the presence of oxygen or of air. The capacity to take up CO or NO is the same as for O_2, *i.e.*, 1 gm. hemoglobin can take up 1.34 cc. of any of these gases.

[1] This figure was given by Huffner (1894). Recently Bernhard and Skeggs (*J. Biol. Chem.*, 147, 19, 1934) have found that 1 gm. of hemoglobin can take up 1.36 cc. of oxygen.

Hemoglobin concentration. There are four methods for determining the hemoglobin concentration: (*a*) spectrophotometry; (*b*) measurement of the respiratory capacity; (*c*) estimation of the iron content; and (*d*) colorimetry. The three first mentioned are considered as standard methods; nevertheless there is a small proportion of hemoglobin that does not combine with oxygen, and in some pathologic cases there are slight variations in the iron content of the blood. Normally there are very small amounts of methemoglobin, carboxyhemoglobin, and inactive hemoglobin.[1]

Colorimetric methods are more frequently used in medical practice, because they are simple and sufficiently accurate. Hemoglobinometers are not always well standardized, and it is necessary to control them before using them. The practice of giving the results as a percentage of an arbitrary normal standard should be abandoned in favor of giving the amount of hemoglobin, in grams, in 100 cc. of blood. There are various reasons for this:

1. The figure considered as normal and given the value of 100 varies in different countries, and even in different parts of the same country.
2. The figure 100 corresponds to different amounts of hemoglobin in the different hemoglobinometers; thus the Haldane standard is 13.4 gm. per cent, Sahli's is 17.3 gm. per cent; the German Society of Internal Medicine gives 16 gm. per cent.

Hemoglobin concentration in 90 per cent of healthy men varies between 14 gm. and 18 gm. per 100 cc. of blood, and between 12 gm. and 15.5 gm. per 100 cc. in women.[2] During the last months of pregnancy there is a decrease in erythrocyte concentration and therefore in hemoglobin concentration, with a return to normal after delivery. In children the hemoglobin concentration is lower than in adults (Table 3). Violent exercise can produce a transitory decrease of as much as 1 gm. per cent, owing to hemolysis (Orías). Persons on a low economic level have less hemoglobin than the wealthy, probably because of deficiencies in nutrition which cause an insufficient production of hemoglobin.

[1] According to van Slyke *et al.* (*J. Biol. Chem.*, **166**, 121, 1946), 0.4 per cent of total hemoglobin is in the form of methemoglobin, and 1.3 per cent is inactive, as it does not take up O_2 or CO.

[2] The following figures were found in the city of Buenos Aires: young men, 15.3 gm. per cent; young women, 13.38 gm. per cent (Orías, O., *Rev. Soc. argent. de biol.*, **10**, 411, 1930).

There is 34 gm. of hemoglobin per 100 cc. of erythrocytes. In Chap. 3 the absolute and relative values of hemoglobin have been considered.

Carbon monoxide intoxication. Carbon monoxide frequently causes intoxication and death owing to accidents, occupational risks, or suicide. Cases of chronic intoxication have also been observed. This gas is formed by the incomplete combustion of carbon in (*a*) stoves, furnaces, and ovens; (*b*) illuminating gas; (*c*) fires; (*d*) internal combustion engines (automobiles, airplanes); (*e*) gunpowder explosions, etc. There is danger when the gas is allowed to accumulate in small, closed, or badly aired premises.

Death is due to asphyxia because CO displaces oxygen from the hemoglobin in the blood, as its affinity for hemoglobin is 300 times that of oxygen, and carboxyhemoglobin is not an oxygen carrier.[1] When 30 per cent of the hemoglobin is saturated with CO there is headache and vomiting; a 50 per cent saturation endangers life; and a 60 to 70 per cent saturation causes severe intoxication and death. The danger to life depends on the concentration of CO in the air breathed and the time the subject is exposed. Thus a 0.01 per cent concentration of CO is well tolerated during many hours; a 0.05 per cent concentration produces signs of intoxication after a few hours; a 0.1 per cent concentration is dangerous after an exposure of one hour; and a 1 per cent concentration kills in a few minutes.[2]

[1] The dissociation curve of hemoglobin in blood intoxicated by CO is altered by shifting to the left, *i.e.*, HbO_2 will give up oxygen only when the oxygen tension is low (see Chap. 28).

[2] The danger of CO poisoning can be calculated, for the practical purposes of industrial hygiene, by multiplying the concentration of CO per million in the air by the hours of exposure. When the figure is 600 only slight symptoms are observed; at 900 there is headache and vomiting; at 1,500 there is danger of death. A popular way of recognizing the presence of CO in the air, used by miners, is to take a bird or a mouse into the mine; these animals rapidly succumb to CO intoxication. Automatic indicators of CO have been manufactured. They contain Hopcalite, a catalyzer which heats up when CO is oxidated. The heat produced warms a thermocouple connected with an indicator, which gives directly the concentration of CO in the air.

The blood and tissues of subjects intoxicated with CO have a typical scarlet color.

Carbon monoxide in the blood can be estimated by chemical methods or by spectrophotometry. There are two characteristic absorption bands in the spectrum of carboxyhemoglobin (Fig. 10). Carboxyhemoglobin is not easily reduced by reagents that reduce oxyhemoglobin, a fact that has been used to distinguish the former from the latter.[1]

Carbon monoxide can be displaced from hemoglobin by oxygen, if the concentration of the latter is greater, according to the mass law: one part of CO is the equivalent of 210 to 300 parts of O_2.

Carboxyhemoglobin is not an absolutely stable compound; otherwise it would be impossible to treat carbon monoxide poisoning with the usual methods. A person intoxicated by CO should be immediately removed from the contaminated atmosphere. If possible, he should be made to breath oxygen or, better still, a mixture of oxygen and 7 per cent CO_2, which increases pulmonary ventilation and thus accelerates the elimination of CO. Artificial respiration should be kept up until the patient breathes normally. Carboxyhemoglobin can be removed from the blood by bleeding and the hemoglobin renewed with the transfusion of normal blood. An intoxicated person recovers gradually when adequate treatment is given, but if the intoxication has been sufficiently severe, sequelae such as pneumonia and lesions (softening) in the central nervous system can occur.

THE STRUCTURE OF HEMOGLOBIN AND HEMOGLOBIN DERIVATES

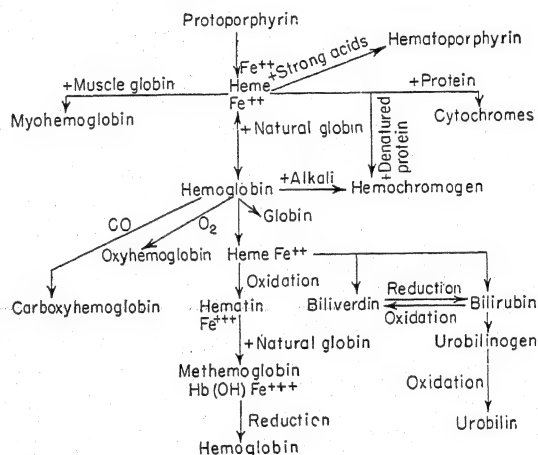
It is possible to separate the two components of hemoglobin, heme and natural globin. Hemoglobin owes its color to heme, which is also responsible for its respiratory function. Heme contains ferrous iron (Fe^{++}) and because of this can take up and set free oxygen in the process of oxygenation and reduction of hemoglobin. This reversible process (hemoglobin $\rightleftarrows$ oxyhemoglobin) is called "oxygenation" to distinguish it from the irreversible oxidation of hemoglobin. When irreversibly oxidized heme is converted into hematin, and hemoglobin into methemoglobin; in these compounds there is ferric iron (Fe^{+++}) and they cannot give up oxygen to the tissues for purposes of respiration.

[1] Consult technical manuals.

Protoporphyrin Compounds

The following are the most important compounds of protoporphyrin:

With ferrous iron (Fe^{++}):
- Heme, or reduced hematin
- Hemoglobin, heme + globin
- Oxyhemoglobin, heme + globin + O_2
- Hemochromogen, heme + denatured globin

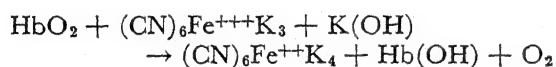
With ferric iron (Fe^{+++}):
- Hematin, oxidized heme
- Hemin, hematin chlorhydrate
- Methemoglobin, hematin + globin + OH
- Cyanhemoglobin, hematin + globin + CN

Hemochromogens. Compounds of heme and nitrogenous bases, *e.g.*, proteins except natural globins, are called "hemochromogens" (Anson and Mirsky). At first this name was given to the compound obtained by treating hemoglobin with an alkali, or oxyhemoglobin with an alkali and a reducing agent. It is a compound of heme with denatured globin, the molecular weight of which is only 16,800.

Hematin. This is an oxidized heme, in which the ferrous iron (Fe^{++}) has been changed to ferric iron (Fe^{+++}). It can be easily prepared by adding dilute acid to hemoglobin or to oxyhemoglobin; the solution takes on a dark brown color. This reaction is used in several hemoglobinometers.

Hemin. This is hematin chlorhydrate prepared by heating blood with glacial acetic acid and a very small amount of NaCl. It forms crystals known as Teichmann's crystals, which are used in legal medicine to identify blood stains (Fig. 9).

Hemin has been obtained by synthesis,[1] and by uniting it to natural globin, hemoglobin has been synthesized.[2] Globin has not yet been prepared synthetically.

Methemoglobin. This is a compound of hematin (oxidized heme) and natural globin. Like other ferric (Fe^{+++}) compounds of protoporphyrin, it does not set free oxygen in a vacuum, and therefore does not give it up to the tissues. It is usually symbolized as Hb(OH), and can be produced *in vivo* and *in vitro* by the action of nitrites, chlorates, anilin, nitrobenzene, etc. It is formed by the action of potassium ferricyanide on oxyhemoglobin, setting free oxygen. This reaction is used for the estimation of oxyhemoglobin:

$$HbO_2 + (CN)_6Fe^{+++}K_3 + K(OH) \rightarrow (CN)_6Fe^{++}K_4 + Hb(OH) + O_2$$

Methemoglobin is a chocolate brown color; it crystallizes, and its absorption spectrum has a typical band in the red, between the C and D lines, the maximum intensity being at 6302 A. On reduction with sodium hydrosulfite, it is transformed into hemoglobin, which can in turn be oxygenated into oxyhemoglobin. When there is marked methemoglobinemia, the subjects have a permanent blue color (cyanosis) of their skin. Enterogenous cyanosis is due to methemoglobin formed by nitrites originated in the intestine. Methemoglobin is slowly disintegrated in the organism by an enzymatic process, which can be accelerated by glucose and methylene blue (M. Brooks).

Methemoglobin combines with CNH and forms cyanhemoglobin, or rather cyanmethemoglobin. In the treatment of cyanide intoxication, moderate doses of nitrites are given by inhalation of amyl nitrite or intravenous injection of sodium nitrite. The nitrites transform part of the hemoglobin or oxyhemoglobin into methemoglobin, which then combines with the cyanide in the plasma. The concentration of cyanide is thus diminished, and the condition of the patient improves. A still better treatment (Hug) consists in the intravenous injection first of sodium nitrite, which forms methemoglobin, and then of sodium thiosulfate, which is an antidote of cyanide,[3] transforming the latter into the less toxic sulfocyanide.

[1] FISCHER and ZEILE, *Ann. d. Chem.*, **468**, 98, 1929.

[2] HILL, R., and H. F. HOLDEN, *Biochem. J.*, **20**, 1326, 1926.

[3] The antidotic action of thiosulfate on cyanide was discovered by Lang, and later studied by Buzzo.

Catahemoglobin is an oxidated hemochromogen, formed by the union of hematin (oxidized heme) with denatured protein.

PORPHYRINS

Porphyrins form part of the active nucleus of chlorophylls a and b, hemoglobin, myohemoglobin, catalase, peroxidase, and cytochromes.

Glycine is a precursor of heme (Shemin and Rittenberg, 1946) and its derivatives (stercobilin) and other porphyrins (uroporphyrin, coproporphyrin). This has been demonstrated by administering glycine with labeled N (N^{15}) or labeled C, which are traced to the pyrrole nuclei of heme. Acetate is also a precursor of part of the pyrrole molecule.

The following porphyrins have been identified so far in different substances of human origin: protoporphyrin 9, mesoporphyrin 9, deutoporphyrin 9, coproporphyrin I and III, and uroporphyrin I. Porphyrin derives from porphin by substitution of the β hydrogens. Porphin is made up of four pyrrole groups joined by four methinic bridges. Fischer has proposed the following terminology for this group of substances: the pyrrole nuclei are numbered with the Roman numerals I to IV; the β hydrogens of the four nuclei are numbered from 1 to 8 with Arabic numerals; the four CH bridges are named α, β, γ, and δ (formula I). The β hydrogens can be replaced by the following groups: methyl (CH_3), ethyl (C_2H_5), vinyl ($CHCH_2$), a propionic acid residue ($CH_2 \cdot CH_2COOH$), etc. From the number of substituted groups, the number of isomers can be calculated. The different isomeric types are called I, II, III, and IV when the substituted groups have the same relative position as the methyl and ethyl groups of etioporphyrins I, II, III, and IV.

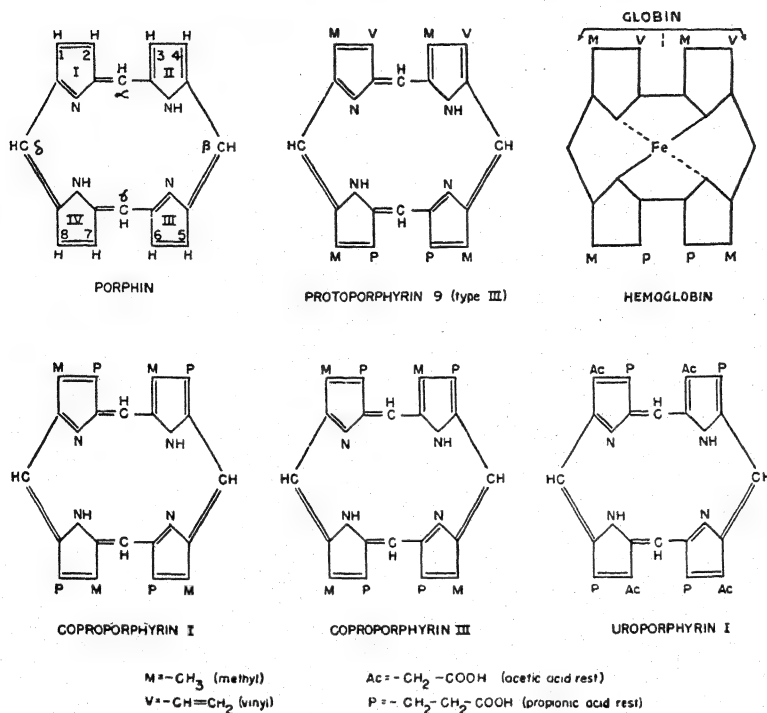
The porphyrin in hemoglobin is protoporphyrin 9 (isomer III). Normally it is found free in small quantities in the erythrocytes and in Harder's gland in rats. Pathologically it can be found in the urine and feces. There is a small amount of free coproporphyrin III in erythrocytes; apparently it is a precursor of protoporphyrin (Watson).

Hematoporphyrin is a laboratory product obtained by the action of strong acids on hemoglobin. In spite of its name it is not the porphyrin in hemoglobin and has never been found in normal or pathologic human tissues or secretions.

A small amount of coproporphyrin is nor-

PORPHIN PROTOPORPHYRIN 9 (type III) HEMOGLOBIN

COPROPORPHYRIN I COPROPORPHYRIN III UROPORPHYRIN I

M = $-CH_3$ (methyl)
V = $-CH{=}CH_2$ (vinyl)
Ac = $-CH_2-COOH$ (acetic acid rest)
P = $-CH_2-CH_2-COOH$ (propionic acid rest)

mally found in the feces (200 to 250 μg daily) and in urine (30 to 80 μg). This is a mixture of coproporphyrins I and III.

The protoporphyrin of the erythrocytes (20 to 40 μg per 100 ml.) increases when there is an intense erythroblastic activity, a deficiency in the utilization of iron, or decomposition of hemoglobin within the erythrocyte.

Porphyrinuria. In certain pathologic conditions the excretion of coproporphyrin in the urine increases; this is known as porphyrinuria. Coproporphyrin I increases when erythropoiesis is abnormally active and in hepatic diseases. Coproporphyrin III is excreted in large amounts in lead poisoning and other intoxications and when there is an infectious condition.

The main porphyrins that have so far been found in man are protoporphyrin 9, mesoporphyrin 9, deuteroporphyrin 9, coproporphyrins I and III, and uroporphyrin I.

Porphyria. This is a disturbance in pyrrole metabolism. There are two main types: (*a*) *congenital porphyria;* (*b*) *acute intermittent porphyria,* including toxic porphyrias produced by certain drugs.

Congenital porphyria is more frequent in males than in females. It appears early in life and has the following signs:

1. Cutaneous lesions due to photosensitivity of the skin provoked by porphyrin.[1] They are situated in areas exposed to light. Reddening and blistering of the skin occur; in extreme cases there is necrosis and mutilation of the extremities.
2. Reddish-brown urine, which becomes red when exposed to the air. It contains large amounts of uroporphyrin I, coproporphyrin I, and other porphyrins. In many cases it has been shown to be transmitted genetically as a recessive character.

Acute intermittent porphyria is more frequently observed in women than in men and is transmitted by heredity. It has the following signs and symptoms:

1. Abdominal pain, constipation, and other digestive disturbances.
2. Muscular weakness, paralysis, and other nervous symptoms.
3. Mental disorders.
4. Usually, dark brown urine, which concontains a mixture of uroporphyrin I with

[1] A subject injected with porphyrins or one who produces porphyrins has no cutaneous alterations if kept in the dark, but the alterations appear when the skin is exposed to light.

other porphyrins. There is also a colorless porphyrinogen.

Bilirubin

Bilirubin is the pigment in human bile; it is also found in blood plasma, which owes its yellow color mainly to this pigment. Normal plasma contains 0.6 (0.4 to 0.8) mg. per 100 cc. It originates in the heme of hemoglobin. Injection of hemoglobin, or of heme, is rapidly followed by an increase in the excretion of bilirubin in the bile. This excretion is also increased when there is intravascular hemolysis. Coupled oxidation of hemoglobin or hematin and ascorbic acid *in vitro* gives rise to green hemochromogen, which is converted to biliverdin; this is then transformed into bilirubin (Lemberg and Legge, 1949). Part of the bilirubin formed (at least 11 per cent) apparently does not derive from hemoglobin (London *et al.*, 1950). Bilirubin is formed partly in the liver and partly outside it.

The extrahepatic origin of bilirubin. Local formation of bilirubin in the tissues has been suspected for many years because of several facts:

1. A yellow color appears where there has been a local extravasation of blood, *e.g.*, bruising of the skin.
2. A long time after blood has been shed into the tissues, crystals are found, which were at first thought to be of a substance called hematoidin, but now have been identified as bilirubin.
3. Jaundice is a prominent symptom in several hepatoses, in which there is extensive degeneration and destruction of liver cells. Therefore the liver must be mainly an organ of excretion of bilirubin, not of its formation.

Mann and Magath gave definite proof of the extrahepatic formation of bilirubin. They removed the liver, or the liver and the intestines, in dogs; 3 to 6 hr. after hepatectomy, the tissues took on a jaundiced aspect, and the concentration of bilirubin in plasma increased.[1] Injection of hemoglobin increased even more the concentration of bilirubin in plasma. The same results have been obtained in eviscerated rats kept alive for 48 hr.

[1] Confirmed by Royer and Cornejo-Saravía (*Compt. rend. Soc. de biol., Paris*, **102**, 424, 1929).

The cells of the reticuloendothelial system, principally in the spleen and bone marrow, are the source of extrahepatic bilirubin. Hemoglobin is taken up by the phagocytes of this system and changed into bilirubin. Mann and his collaborators found more bilirubin in the blood of the splenic vein and in the blood leaving the limbs than in arterial blood. The difference in bilirubin concentration in arterial and venous blood from a limb disappeared when the bones (consequently the bone marrow) were extirpated. Splenectomy alone does not suppress the formation of extrahepatic bilirubin, which continues in other parts of the organism. If the reticuloendothelial system is blocked by the injection of colloid suspensions (*e.g.*, colloidal iron), a transitory decrease in the excretion of bilirubin in the bile occurs.

The hepatic origin of bilirubin. The part played by the liver in the formation of bilirubin has been clearly demonstrated in geese. When they are intoxicated with hydrogen arsenide there is hemolysis, hemoglobinemia, and (later) jaundice. If the liver is removed, jaundice does not occur (Minkowski and Naunyn). The absence of jaundice was attributed to the removal of the hepatic cells, but McNee believes it is due to the suppression by hepatectomy of a large part of the reticuloendothelial system, in this species situated mainly in the liver.

In the dog the injection of toluylenediamine provokes jaundice. Previous removal of the liver reduces jaundice to one-third its intensity (Melchior, Rosenthal, and Licht). From these experiments it has been concluded that in this species two-thirds of the bilirubin is formed in the liver and one-third is of extrahepatic origin.

The double origin of bilirubin, hepatic and extrahepatic, is now universally admitted. Conclusive evidence has not yet been obtained as to (*a*) whether more bilirubin is formed outside the liver (Mann) or in the liver (Rosenthal); (*b*) whether bilirubin is formed in the cells of the reticuloendothelial system (Aschoff-McNee) or mainly in the hepatic cells (Rosenthal).

Direct and indirect bilirubin. Van den Bergh has used diazonium as a reagent for bilirubin.

1. *The direct reaction.* If blood plasma or serum to which the diazoreagent is added gives a violet color in 10 to 30 sec., there is said to be a direct reaction (van den Bergh). This direct reaction is given by

bilirubin which has passed through the hepatic cells.[1] It is obtained with bile; also with blood plasma in those cases of jaundice due to the reabsorption of bile because of an obstacle in the bile ducts, or because of dislocation of the hepatic cells in the course of degenerative or inflammatory processes of the liver.

2. *The indirect reaction.* In this reaction, observed in hemolytic jaundice, alcohol must first be added to the plasma or serum, and then the diazoreagent. The indirect reaction is supposed to be given by bilirubin formed in the reticuloendothelial system without passing through the hepatic cells.

Other differences between direct and indirect bilirubin have been noted: (*a*) direct bilirubin passes through the kidney into the urine; (*b*) indirect bilirubin can be extracted with chloroform. This last is not a definite proof that indirect bilirubin is a different substance, as the extraction is conditioned by the pH in aqueous solutions (Kerppola and Leikola). In plasma the variations in pH do not modify the reaction to any extent, but an increase in phosphatides favors the extraction (López-García and Zelasco). The cause of the direct and indirect reaction is not known. It is supposed that in the first case bilirubin is only absorbed by the serum albumin, while in the second case it forms a compound from which it must be separated by alcohol before it can react with diazonium (Coolidge).

A yellow reaction with diazonium is given by plasma which has a high concentration of urobilinogen (Varela-Fuentes). In patients with jaundice due to obstruction of the bile ducts by gallstones, cancer, or any other obstacle, the serum gives up bilirubin when extracted with ether (ether-soluble bilirubin).

Jaundice. When there is an abnormally high concentration of bilirubin in the plasma, the skin and mucosa take on a yellow color (jaundice) and the urine becomes yellow. Jaundice has been classified into (*a*) *prehepatic jaundice* of hemolytic and other origins (among the hemolytic varieties there is *hemolytic jaundice*, with increased fragility of the erythrocytes); (*b*) *hepatic jaundice*, due to lesions in the hepatic cells or canals caused by intoxication or infection (the vulnerable point of the bile canals is the ampulla where the intercellular spaces join the ducts that have a differentiated wall [Eppinger]); (*c*) *obstructive jaundice*, also called *posthepatic jaundice* (Ducci), which is due to reabsorption of bilirubin into the blood and lymph because of an obstacle in the bile ducts.

[1] According to the classical conception extrahepatic bilirubin is taken up by the Kupffer cells and then passed to the liver cells, which modify it and excrete it into the bile ducts.

In certain species, *e.g.*, the toad, *Bufo arenarum*, bile contains only biliverdin, which gives an olive green color to the plasma when it accumulates in the blood (Cabello, 1943); "chloricia" would be the appropriate name for this type of jaundice. In mammals biliverdin has never been found in the plasma.

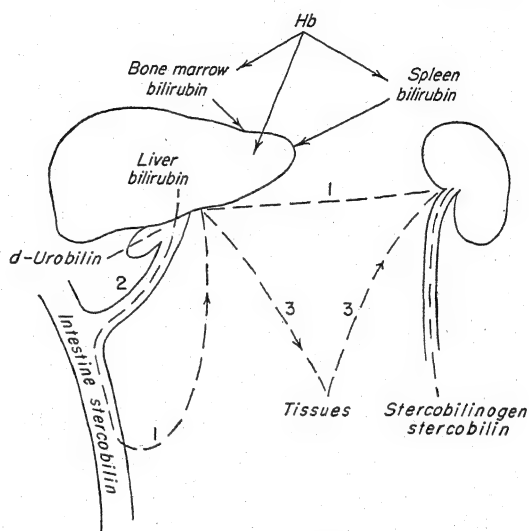



FIG. 11. Origin and cycle of urobilin. 1, normal route, intestine-liver-blood-kidney; 2, origin in infected gall bladder; 3, small quantity to and from the tissues.

STERCOBILIN AND UROBILIN

When bile reaches the intestine, bilirubin is converted into stercobilinogen by the action of the intestinal bacteria. Stercobilinogen when oxidized gives stercobilin. Watson has obtained crystallized stercobilin from feces (1932) and urine (1933).

Bilirubin is converted by a process of reduction into mesobilirubinogen, which may then follow one of three main paths (Watson):

1. It is converted into L-stercobilinogen by the simultaneous action of bile and feces; L-stercobilinogen is then oxidized into L-stercobilin.
2. If there is no bile, it is converted by the feces into urobilin IXα. This can also be done *in vitro* by a process of reduction (Siedel, 1936).
3. In bile contaminated by bacteria it is converted into D-urobilin. Stercobilin and the different urobilins can be identified by the green fluorescence they give out in the pres-

ence of zinc salts (Jaffé, 1868). Stercobilinogen and urobilinogen in the presence of *p*-dimethylaminobenzaldehyde give a red color; this reaction (Ehrlich's reaction) is used for quantitative determination of these substances.

The amount of stercobilinogen and stercobilin in the feces is dependent mainly on the amount of bilirubin that reaches the intestine. If bile is prevented from arriving into the intestine, there is no bilirubin or stercobilinogen in the feces, and no urobilin in the urine.

Certain bacteria are of fundamental importance in the formation of stercobilinogen. Thus, if aureomycin is administered in doses such that the intestinal contents become free from coli bacilli, stercobilinogen disappears from the feces, bile, and urine.

At least 15 to 30 per cent of stercobilin does not come from hemoglobin in the erythrocytes. When glycine labeled with N^{15} is given, it can be traced to stercobilin almost immediately and not only after an interval of 100 to 125 days, as would be the case if it were first incorporated into hemoglobin and appeared in stercobilinogen only after disintegration of the erythrocytes (London *et al.*, 1949). It seems that in pernicious anemia a larger proportion of stercobilinogen is originated in this way.

Intestinal stercobilinogen and stercobilin[1] follow two paths: (*a*) up to 85 per cent is partially disintegrated and excreted in the feces (40 to 280 mg./day, measured as stercobilinogen); (*b*) the rest is reabsorbed by the intestinal mucosa and taken by the portal blood to the liver (Fig. 11). If the liver is functioning normally most of the stercobilin is retained, and its structure is modified by chemical processes not yet well understood. A very small amount that is not retained by the liver passes into the blood and is distributed throughout the body. Its presence in the blood has not been demonstrated by analytic methods, but there is indirect evidence of its existence. In diseased persons it has been found in the blood and its amount measured. It is finally excreted by the kidney as stercobilinogen (urobilinogen), which by oxidation is converted into stercobilin (urobilin). The total amount eliminated daily is not more than 0.6 mg. Very small amounts have been found in tissues (Royer).[1]

[1] This term will cover stercobilin and stercobilinogen. The latter is more abundant; from 50 to 200 mg. is excreted daily in human feces (Royer).

Müller formulated in 1892 the so-called "enterohepatic theory," *i.e.*, that stercobilinogen (urobilinogen) is formed in the intestine and reabsorbed in part. Most of this is retained by the liver, the remainder being eliminated by the kidney. For this reason urobilin diminishes and finally disappears from the urine of patients with an obstruction in the common bile duct; it reappears if these patients are given bile by mouth. Experimental proof of this theory is now available.

The biliary origin of urobilin can be demonstrated by making a biliary fistula in a dog, so that bilirubin cannot pass into the intestine. After a few days stercobilin disappears from the feces and urobilin from the urine, the tissues, and the bile (Royer and Cornejo-Saravía). The introduction of bile or bilirubin into the intestine is followed by the reappearance of stercobilin in the feces and urobilin in the urine.

The liver plays an important part in retaining and regulating the excretion of these substances. A few hours after hepatectomy, stercobilin can be found in the blood plasma, where it increases progressively. This is due to its absorption from the intestine and the absence of retention by the liver. The intestinal origin can be easily demonstrated: a biliary fistula is made and hepatectomy is not performed until the feces contain no more stercobilin; in this case there is no accumulation of stercobilin in the blood. If bile or stercobilin is introduced into the intestine, it appears in the blood (Royer and Cornejo-Saravía).

The capacity of the liver to retain stercobilin can be demonstrated by injecting this substance into the portal vein and estimating it in the blood of the suprahepatic veins. This capacity diminishes considerably when the liver is damaged by some toxic substances, *e.g.*, chloroform (Royer).

An increase in urobilinogen and urobilin in the urine is a sign of hepatic insufficiency (Royer, Watson, López-García). Urobilinuria also increases when there is an excessive destruction of erythrocytes (pernicious anemia, injection of hemolytic substances). In an infected gall bladder, part of the bilirubin of the bile is transformed into urobilin by bacterial

[1] Jaffé discovered urobilin in 1869. Urobilinogen was identified as mesobilirubinogen by Fischer and Meyer-Betz (1917). Crystallized stercobilin was prepared by Watson from feces (1933) and urine (1934). Seiden and Meyer obtained urobilin IXα by synthesis.

activity (Royer, López-García), and the latter increases both absolutely and relatively to bilirubin.

BIBLIOGRAPHY

Hemoglobin

ANSON, M. L., and A. E. MIRSKY, *Physiol. Rev.*, **10**, 506, 1930 (hemoglobin and heme).

BARCROFT, J., "The Respiratory Function of the Blood," Part II, Hemoglobin, Cambridge, New York, 1928.

DRABKIN, D. L., *Physiol. Rev.*, **31**, 345, 1951 (metabolism of chromoproteins).

LILIENTHAL, J. L., Carbon Monoxide, *Pharmacol. Rev.*, **2**, 324, 1950.

PAULING, L., *et al.*, *Science*, **110**, 543, 1949 (hemoglobin in drepanocytes).

ROUGHTON, F. J. W., and J. C. KENDREW, "Hemoglobin" (symposium in memory of J. Barcroft, Cambridge, June, 1948), Interscience Publishers, Inc., New York, 1949.

Myohemoglobin

BIOREK, G., Myohemoglobin, *Acta med. Scandinav.*, **133** Suppl. 226, 1949.

MILLIKAN, G. A., *Physiol. Rev.*, **19**, 503, 1939.

THEORELL, H., and C. DE DUVE, *Arch. Biochem.*, **12**, 113, 1947.

Porphyrins

DOBRINER, K., and C. P. RHOADS, *Physiol. Rev.*, **20**, 416 1940.

FISCHER, H., and H. ORTH, "Die Chemie des Pyrrols," Akademische Verlagsgesellschaft m.b.H., Leipzig, 1937.

RIMINGTON, C., *Lancet*, **2**, 551, 1951.

SHEMIN, D., and D. RITTENBERG, *J. Biol. Chem.*, **166**, 621, 1946.

SHEMIN, D., and J. WITTENBERG, "Ciba Foundation Conference on Isotopes in Biochemistry," J. & A. Churchill, London, 1951.

WATSON, C. J., *Blood*, **1**, 99, 1946.

WATSON, C. J., *Arch. Int. Med.*, **86**, 792, 1950.

WATSON, C. J., and E. A. LARSON, *Physiol. Rev.*, **27**, 478, 1947.

Bilirubin

BERGH, H. VAN DEN, "Der Gallenfarbstoff in Bluete," Barth, Leipzig, 1928.

LEMBERG, F., and J. W. LEGGE, "Hematin Compounds and Bile Pigments," Interscience Publishers, Inc., New York, 1949.

MALLOY, H., and K. A. EVELYN, *J. Biol. Chem.*, **119**, 481, 1937 (methods of determination).

MANN, F. C., and J. L. BOLLMAN, *J. A. M. A.*, **104**, 371, 1935 (origin of jaundice).

ROSENTHAL, L., H. LICHT, and E. MELCHIOR, *Klin. Wchnschr.*, No. 44, 2076, 1927 (origin of jaundice).

Urobilin

FISCHER, H., and H. ORTH, "Die Chemie des Pyrrols," Akademische Verlagsgesellschaft m.b.H., Leipzig, 1937.

LONDON, I. M., *et al.*, *J. Biol. Chem.*, **184**, 351, 359, 365, 1950.

ROYER, M., "La Urobilina en el Estado Normal y Patológico," 2d ed., El Ateneo, Buenos Aires, 1942.

WATSON, C. J., *Arch. Int. Med.*, **19**, 196 and 206, 1937.

WATSON, C. J., *Proc. Soc. Exper. Biol. & Med.*, **49**, 641, 1942.

WATSON, C. J., in H. Downey, "Handbook of Hematology," vol. 4, Hoeber, New York, 1938, p. 2447.

WATSON, C. J., *Harvey Lect.*, **44**, 411, 1948.

CHAPTER 5

The Life of the Erythrocytes

THE FORMATION OF the blood cells is known as hemopoiesis. Erythropoiesis is the formation of the red blood cells, leukopoiesis that of the white blood cells, and granulopoiesis that of the polymorphonuclear granulocytes.

EMBRYONIC AND FETAL ERYTHROPOIESIS

Three stages can be distinguished in the development of the erythrocytes in the embryo and fetus: (*a*) mesoblastic; (*b*) hepatic; (*c*) myeloid. Each one of these stages is predominant during 3 months of intrauterine life, and the passage from one to another takes place gradually.

The mesoblastic period begins in the early stages of embryonic development. All the blood cells originate in the totipotential embryonic mesenchyma. Vascular islets (Wolff and Pander) are formed in the extraembryonic structures, on the mesenchyma of the yolk sac; blood vessels and erythrocytes are formed in these islets. The vascular areas join up with each other to form a network, which is later connected with the chorion and with the heart. The islets of Wolff and Pander are at first compact columns of cells but later the peripheral cells flatten out and become primitive endothelial cells, while those in the center of the column take on a rounded shape and are separated from each other by a primitive plasma. These free cells (erythroblasts) are large nucleated cells. If they become prematurely loaded with hemoglobin, they give rise to large nucleated erythrocytes. Primitive erythrocytes are also formed by proliferation of the endothelial cells, at first in the extraembryonic vascular areas and later within the embryo. Finally the circulating cells also multiply and form primitive blood cells. This type of development is seen between the third and sixth week of embryonic life, begins to diminish in the second month, and has ceased completely by the fourth month.

The hepatic stage commences in the course of the second month of embryonic life, but it is not fully developed until the third month, when the liver becomes the principal organ of erythropoiesis; later the spleen also takes part in this process. The liver produces normocytes, *i.e.*, nonnucleated red cells of normal size. From the third month onward, only 8 per cent of the erythrocytes have a nucleus.

In the myeloid stage erythropoiesis takes place mainly in the bone marrow. This stage begins during the second month and develops in the course of the fifth month, and as the liver loses importance the bone marrow gains it. During the last three months of fetal life, the bone marrow becomes the principal site of erythropoiesis.

ERYTHROPOIESIS IN THE ADULT

After birth, in man, erythropoiesis is carried out exclusively in the bone marrow; in normal conditions no erythrocytes are formed in the liver and the spleen.

There are two theories as to the origin of the blood cells: the unicist (monophyletic) and the pluralist (polyphyletic) theories. The former maintains that the mesenchyma gives rise to a primitive blood cell, called hemocytoblast (Ferrata), lymphoidocyte (Pappenheim), myeloblast (Naegeli), etc., from which all the blood cells arise. There are two main pluralist theories:

1. The dualist theory postulates the existence of a myeloid series of cells produced in the bone marrow, in which a primitive cell gives rise to the erythrocytes and the granulocytes; a second series of lymphoid cells is produced in the lymphatic organs and gives rise to the lymphocytes and monocytes.
2. Schilling has proposed a theory in which the blood cells have three different origins: a

myeloid series of cells produces erythrocytes and granulocytes; lymphocytes and monocytes have each a separate source.

Table 5 gives the principal stages in the development of the different adult blood cells.[1]

Table 5. Hemocytopoiesis in Adult Man

Mesenchymatic cell in hemopoietic tissues

Lymphoid organs	Reticuloendothelial system	Myeloid tissue			
		Myeloblast	Megakaryoblast	Proerythroblast	Promegaloblast†
Lymphoblasts				Basophil erythroblast	Basophil megaloblast
	Monoblast	Promyelocyte			
Prolymphocyte		Myelocyte		Polychromatophil erythroblast	Polychromatophil megaloblast
Lymphocyte	Monocyte	Metamyelocyte	Megakaryocyte	Orthochromatic normoblast	Orthochromatic megaloblast
		Granulocyte*	Thrombocytes	Erythrocyte	Macrocyte

* There are three series of myelocytes: neutrophils, eosinophils, and basophils, which give rise to the corresponding type of granulocyte.

† This series appears only when there is insufficiency of the hematinic or antianemic principle, which is necessary for the normal maturation of the erythroblasts.

In the development of the erythrocytes several stages can be differentiated. The primitive hemocytoblast gives rise to a proerythroblast,[2] the precursor of the erythroblasts. This is a nucleated cell with a strongly basophil protoplasm, which gradually loses its basophilia, and after passing through a polychromatophil stage, becomes acidophil (orthochromatic) like the adult erythrocyte and is called a normoblast. The nucleus shrinks and then disappears by extrusion from the cell, although some observers maintain that it is dissolved (karyolysis). In certain pathologic conditions the nucleus breaks up into fragments (karyorexis) and then disappears.

[1] The sum of the circulating erythrocytes and the erythroid cells of the bone marrow has been called the "erythron" (Boycott) or "erythrocytic organ." Its total volume is about 1,500 cc. in the adult.

[2] Sabin and Doan have given the name of "megaloblast" to the proerythroblast. Previously Ehrlich had given this name to a cell found only in pathologic cases (pernicious anemia, lack of erythroblastic maturation factor); this is a large cell with a nucleus of typical structure, which has a delicate chromatin network. The name of megaloblast should be given only to this abnormal cell.

Reticulocytes. Young erythrocytes still have some basophilic substance, which is evident either as polychromasia or in the form of a fine basophilic reticulation of the cytoplasm (Cesàris Demel); these cells are known as reticulocytes. The reticulation can be revealed by supravital staining of the fresh erythrocytes with cresyl blue, which precipitates the basophilic substance; if a blood smear is then made, it can be further stained with the usual dyes.

Normally 0.5 to 1.5 per cent of the erythrocytes are reticulocytes. The proportion is greater when erythropoiesis is stimulated, *e.g.*, during the period of blood regeneration after hemorrhage, in the polycythemia of high altitude, and in hemolytic jaundice. The increase in reticulocytes is a valuable sign of improvement in pernicious anemia, and serves as a guide in the treatment of this disease with liver extracts, as it precedes the increase in erythrocytes (Fig. 14).

Bone marrow. The total amount of bone marrow in the body weighs from 1.5 to 3.5 kg. It has the following functions:

1. The formation of erythrocytes (erythropoiesis).

2. The formation of granulocytes (granulopoiesis).
3. The formation of thrombocytes.
4. The destruction of erythrocytes.

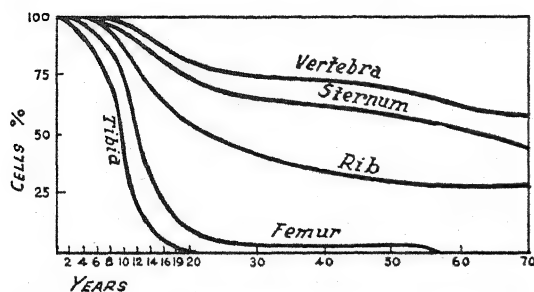



FIG. 12. Activity of bone marrow at different ages. The curves of the shaft of the fibula, radius, and ulna are similar to that of the tibia; that of the humerus is similar to that of the femur.

The bone marrow can develop a considerable erythropoietic activity, and in two weeks it can produce several times its weight of erythrocytes (Whipple). In infancy and when there is an increased erythropoietic activity, the bone marrow has a red color. From puberty onward the marrow in the shafts of the long bones of the limbs becomes loaded with fat, takes on a yellow color, and has little or no erythropoietic activity; this process is completed by the twentieth year of life Hemopoiesis continues throughout life in the marrow of the vertebrae, sternum, ribs, pelvis, and bones of the skull (Fig. 12). The hemopoietic activity of the bone marrow can be examined by making a smear of a sample obtained by puncture of the sternum.[1]

There is still doubt as to whether the erythrocytes are formed outside the blood vessels or in closed sinusoids.

THE REGULATION OF ERYTHROPOIESIS AND THE FORMATION OF HEMOGLOBIN

The most important quantitative data on the erythrocytes are

1. The concentration, or number of erythrocytes per cubic millimeter.
2. The total amount in all the blood in the body, given in liters or in cubic centimeters.
3. The relative volume in 100 cc. of blood, measured by the hematocrit.

Variations in the concentration of erythrocytes can be due to

[1] Pianese, 1903; Ghedini, 1908; Seifarth, 1923.

1. Hemoconcentration produced by loss of water from the plasma (profuse sweating, diarrhea, vomiting, polyuria, shock, etc.), or by loss of plasma (burns).
2. Dilution of the blood.
3. Sudden mobilization of the erythrocytes in the spleen and other blood stores.

The total number of circulating erythrocytes is the result of a dynamic equilibrium between their formation and destruction. A temporary alteration of this equilibrium can be brought about by the sudden outpouring of erythrocytes into the circulation by contraction of the spleen, or inversely by the storage of erythrocytes in blood reservoirs. The equilibrium between formation and destruction of erythrocytes is fairly constant. Variations in the number of erythrocytes are seldom greater than 300,000 and exceptionally 500,000 on the same day, and the erythrocyte count of an individual determined on different days gives a fairly constant figure. The loss of erythrocytes by hemorrhage is replaced in a few days or weeks by an increased production until the normal concentration is reached.[1] On the other hand, if erythrocytes are injected so as to increase their concentration, the number in excess is destroyed until the normal concentration is restored.

Formation of the stroma. Apparently there is never a lack of raw materials necessary for the production of the red cell stroma, such as nucleoproteins, globulins, phosphatids, cholesterol, and minerals.

Proteins and the regeneration of hemoglobin. A normal diet must provide a sufficient amount of high-quality protein to maintain a normal concentration of hemoglobin and plasma proteins. Whipple, Hooper, and Robscheit[2] have studied the regeneration of hemoglobin in dogs. The animals were submitted to repeated hemorrhages, so as to diminish the concentration to 30 per cent of the normal value. In the interval between two bleedings the protein under consideration was given in the food or injected, and the amount of hemoglobin regenerated was measured. Hemoglobin regeneration was accelerated by several proteins

[1] In some cases erythropoiesis is stimulated by moderate hemorrhage, and the erythrocyte count is increased above the initial level.

[2] *J. Exper. Med.*, 77, 375, 1943; *Nutrition Rev.*, 1, 284, 1943.

and amino acids.[1] Excellent results were obtained when the following substances were given by mouth: liver, kidney, meat, or amino acids. The same results are obtained by injection of blood plasma, hemoglobin or globin, or the product of digestion of these proteins or of caseinogen. A mixture of the amino acids indispensable to maintain the body weight gives rise to the formation of great amounts of hemoglobin and plasma proteins. On a carbohydrate diet (bread and sugar) there is not much regeneration of hemoglobin. The liver plays a direct or an indirect part in the utilization of food for hemoglobin regeneration. Glycine contributes to make up a large part of the pyrrole constituent of porphyrins and heme.

Iron. Nearly all the iron in the blood is found within the erythrocytes, forming part of the hemoglobin molecule. In 100 gm. of hemoglobin there is 0.34 gm. of Fe, about 50 mg. in 100 cc. of blood, and 2 to 3 gm. in all the blood of an adult; the total amount of iron in the body is from 3 to 4 gm. There is only 0.12 to 0.14 mg. of iron in 100 cc. of plasma.[2] There are iron reserves in the bone marrow, the liver, and the spleen. A crystallized protein containing 20 per cent iron has been obtained from these organs; it is called ferritin. These reserves are important in infancy; they prevent breast- or bottle-fed babies from having a decrease in the hemoglobin concentration of the blood in spite of the fact that milk has very little iron.

The diet should provide 10 to 12 mg. of iron daily, and 15 mg. in the case of pregnant women. Inorganic iron, especially ferrous iron, is well absorbed and utilized, better than organic combinations of iron. The principal factor in the regulation of iron absorption is the amount of iron in the body. If there are abundant reserves, little is absorbed; if they are low, more is absorbed (Whipple *et al.*).[3] Absorption of iron increases during pregnancy and is favored by vitamin C. An excess of alkali, of phytate (brown bread, oats), or of phosphate in the diet diminishes the absorption of iron, because part of it becomes insoluble (see Chap. 45, Mineral Metabolism).

Iron deficiency causes hypochromic microcytic anemia. In this type of anemia, but not in others, iron treatment gives good results.[1] Anemia due to iron deficiency may be caused by a prolonged exclusive milk diet or by chronic hemorrhages. Iron deficiency is an important factor in the anemia produced by hookworm (*Necator* or *Ancylostoma*); usually it can be prevented by a meat diet (Fulleborn) and is improved by iron treatment (Oswaldo Cruz, Jr.).[2]

[1] *J. Exper. Med.*, **81,** 171, 1945; **85,** 243 and 267, 1947.

[2] Powell, J. F., *Quart. J. Med.*, 13 (49), 19, 1944.

[3] Balfour, W. M., *et al.*, *J. Exper. Med.*, **76,** 15, 1942; Hahn, P. F., *et al.*, *J. Exper. Med.*, **78,** 169, 1943.

Copper. White rats fed exclusively on cow's milk suffer from a fatal type of anemia. Iron treatment does not improve the condition of these animals, but if copper is added to the iron treatment, hemoglobin is normally regenerated and the animals are cured.[3] Copper deficiency is unknown in man; if copper is an indispensable element, sufficient amounts are provided in food and water. Copper treatment has not improved the result of other treatments in human anemias. Human serum has 0.11 mg. per cent copper in man and 0.13 mg. per cent in woman.

Oxygen tension. A very important factor that regulates erythropoiesis is the partial pressure of oxygen in the atmosphere. A decrease in oxygen pressure stimulates erythropoiesis; an increase has an inhibitory or moderating effect. The stimulating effect of low oxygen tension was first observed by Viault (1889), who found an increase in the concentration of erythrocytes in the blood of persons living in the high mountains of Peru; this is known as high-altitude polycythemia.[4] The same fact has been observed in the blood of the inhabitants of many cities situated at high altitudes in Peru, Bolivia, and Mexico. The increase in erythrocyte concentration is proportional to the height above sea level (Table 6). There is an increase in the concentration and in the total amount of erythrocytes. In the course of a rapid ascent in an airplane, at first there may be simply a transitory increase in the concentration of erythrocytes, due to contraction of the spleen provoked by anoxia, emotion, or muscular effort, and thus stored erythrocytes are sent into the circulation. When the subject remains at high altitude for some time, there is an increase in the formation

[1] Hahn, P. F., *Medicine*, **16,** 249, 1937; Heath, C. W., and A. J. Pathek, *Medicine*, **16,** 297, 1937.

[2] Cruz, W. O., Jr., and R. P. de Mello, *Mem. Inst. Oswaldo Cruz*, **42,** 401, 1945.

[3] Elvehjem, C. A., *Physiol. Rev.*, **15,** 471, 1935.

[4] The term "hyperglobulia" was used at first, but it should be abandoned as it does not indicate clearly whether there is an increase in the number of the erythrocytes or in their size.

of erythrocytes, preceded by an increase in reticulocytes. There is also an increase in hemoglobin production and in the total amount of hemoglobin in the body. More hemoglobin is, therefore, destroyed, and bilirubin and stercobilin increase (Merino, 1950).

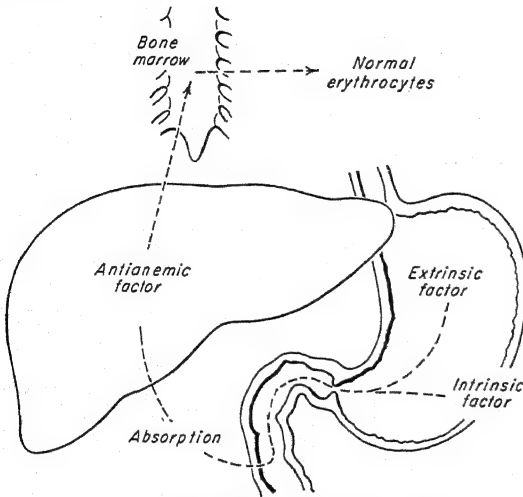



FIG. 13. Antianemic or hematinic factor.

An increase in oxygen tension, produced experimentally by keeping animals in chambers where the atmosphere has been modified by increasing the concentration of oxygen, diminishes the concentration of erythrocytes.

Erythroblastic maturation factor. This factor is also known as the antianemic or hematinic factor. When there is a deficiency of this factor a typical deviation of erythropoiesis occurs; megaloblasts, which develop into megalocytes, are formed instead of normoblasts, which develop into normocytes. This is observed in pernicious anemia and in other macrocytic anemias related to this disease. In pernicious anemia (also called "Addison's" or "Biermer's" anemia) there are megaloblasts in the bone marrow, with production of megalocytes. Therefore, there is anemia, with megalocytes in the circulating blood. Apparently there is also an increased destruction of erythrocytes, as there is an increase of iron and bilirubin concentration in plasma, and an increase in the excretion of stercobilin. There is no hydrochloric acid in the gastric juice (achlorhydria), and after a time severe lesions in the spinal cord are observed. The disease was fatal within a short time until Minot and Murphy (1926) showed that ingestion of liver or kidney in large quantities produced a remission of the disease. Later it was shown that desiccated defatted hog stomach had the same effect. The preparation of very active liver extracts which could be injected has been an important advance in the treatment of pernicious anemia (Gänsslen, 1930).

Castle (1929)[1] observed that if gastric juice and meat are incubated the hematinic principle is formed. This factor would therefore be the result of the interaction of two others, an extrinsic and an intrinsic factor (Fig. 13). The extrinsic factor is found in meat, yeast, etc., and is destroyed by heating. It is vitamin B_{12}. The intrinsic factor is formed in the gastric and duodenal mucosae. By the interaction of the two factors in the stomach and intestine, the antianemic factor is absorbed and stored in the liver, from which it goes to the bone marrow to stimulate normal maturation of erythroblasts and normocytic erythropoiesis.

In pernicious anemia there is deficiency of the intrinsic factor; the extrinsic factor may be found in the intestine. Treatment with liver,

Table 6. Effect of High Altitude on Erythrocyte Concentration

	Altitude		*Erythrocytes, millions per cu. mm.*	*Hemoglobin, gm./100 ml.*	*Reticulocytes, per cent*	*Blood volume, liters*	*Hematocrit*	*Blood bilirubin, mg./100 ml.*	*Leukocytes, thousands per cu. mm.*
	Ft.	*M.*							
Lima	0	0	5.14	16	0.5	5.02	46.8	0.76	6.8
Oroya	12,238	3,730	5.67	18.8	0.8	6.15	54.1	1.47	6.5
Morococha	14,895	4,540	6.15	20.7	1.5	6.98	59.9	1.56	6.9

Source: HURTADO, M., C. MERINO, and E. DELGADO, *Arch. Int. Med.*, **75**, 284, 1945.

[1] CASTLE, W. B., and MINOT, G. R., "Pathological Physiology and Clinical Description of the Anemias," Oxford, New York, 1936; MINOT, G. R., and M. B. STRAUSS, *Vitamins & Hormones*, **1**, 269, 1943.

stomach, or concentrated preparations of the antianemic factor provokes a marked increase in the reticulocytes, which can make up 10 to 40 per cent of the circulating red blood cells; later normal erythrocytes increase, and the normal concentration of red cells is restored in one to two months (Fig. 14). In cases with a less marked anemia, reticulocytosis is also less marked, although the erythrocyte count increases. The potency of an extract is measured by its effects on patients with pernicious anemia.

A deficiency in the antianemic factor can be caused in several ways: (*a*) by the absence of the extrinsic factor in the diet (tropical anemias, pellagra); (*b*) by deficiency in the intrinsic factor (pernicious anemia, gastric cancer, total or subtotal gastrectomy, pellagra); (*c*) by deficient intestinal absorption (chronic enteritis, sprue, pellagra, infestation by *Diphyllobotrium latum*); (*d*) by deficient storage in the liver (extensive chronic lesions of the liver); (*e*) by deficiency in the response of the bone marrow to the antianemic factor (exceptional cases of myxedema).

Folic acid. This vitamin, so called because it was first obtained from spinach leaves (also called folacine, or *Lactobacillus casei* factor), is one of those in the B complex (see Chap. 49). Its deficiency produces in several species an anemia that is cured by treatment with folic acid. It has the biological and chemical properties of pteroylglutamic acid, and synthetic pteroylglutamic acid has its effect. It exerts a powerful therapeutic activity in certain macrocytic anemias caused by deficient diets. It provokes first an increase in reticulocytes, then an increase in erythrocytes and hemoglobin concentration. It does not, however, prevent the development of neurologic disorders seen in pernicious anemia; therefore it should not be used alone in treatment of this disease.

Certain substances chemically related to pteroylglutamic acid, *e.g.*, 4-aminopteroylglutamic acid (aminopterin), are known as folic acid antagonists, because they inhibit the effects of folic acid and provoke symptoms of folic acid deficiency. Folic acid is converted into the citrovorum factor (CF),[1] or folinic acid, by liver tissue, a process which is inhibited by folic acid antagonists.[1] CF can replace folic acid; it has antianemic activity, and is much more potent in the reversal of the effect of folic acid antagonists. CF is the prosthetic group of a coenzyme essential for the transference of formate,[2] a chemical step of importance in the synthesis of purines and thymine in nucleic acids, also of serine, choline, and methionine.[3] Thymine has antianemic activity if given in large doses.

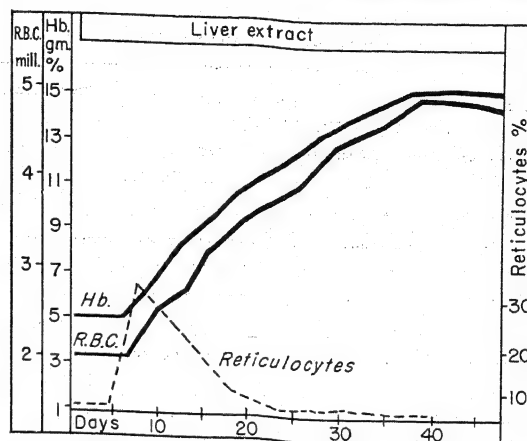



FIG. 14. Effect of liver extract in a case of pernicious anemia.

Vitamin B_{12}. This substance has been found in liver extracts; it forms red crystals which contain cobalt. It was first identified in 1948 by its antianemic properties (Smith *et al.*) and as a growth factor for *Lactobacillus lactis* Dorner (Rickes *et al.*). It is not toxic and exerts its physiologic effects at very small doses; 0.5 to 1 μg is the amount required daily by a normal adult man. Doses of 10 to 150 μg have produced complete remission in cases of pernicious anemia. Reticulocytes increase first, then the erythrocytes and hemoglobin concentration. Megaloblasts and megalocytes disappear from the blood and bone marrow, and the erythrocyte count returns to the normal level. Lesions on the tongue are healed, and neurologic disturbances are either prevented or, when already established, cured if they are of only moderate severity. Vitamin B_{12} is not active when given by mouth

[1] Sauberlich and Baumann (*J. Biol. Chem.*, **176**, 165, 1948) discovered a factor necessary for the growth of *Leuconostoc citrovorum* (hence its name) which was found in liver extracts and in yeast.

[1] BROQUIST, H. P., E. L. R. STOKSTAD, and T. H. JUKES, *J. Biol. Chem.*, **185**, 389, 1950; BURCHENAL, J. H., *et al.*, *Proc. Soc. Exper. Biol. & Med.*, **74**, 735, 1950; **76**, 382, 1951.

[2] PLAUT, G. W. E., *et al.*, *J. Biol. Chem.*, **184**, 795, 1950.

[3] WELCH, A. D., and W. J. SAKAMI, *Federation Proc.*, **9**, 245, 1950; SKIPPER, H. E., *et al.*, *Cancer Research*, **10**, 510, 1950.

unless gastric juice is administered simultaneously,[1] as occurs with Castle's extrinsic factor. It has been found in the feces in cases of pernicious anemia; probably it is not absorbed owing to deficiency of gastric intrinsic factor. Serum from normal subjects or from patients with pernicious anemia treated with B_{12} ripens megaloblasts into normoblasts *in vitro*, but neither vitamin B_{12} nor normal gastric juice separately has this effect. The hemopoietic factor in normal serum is formed by the interaction of vitamin B_{12} (Castle's extrinsic factor) with a gastric intrinsic factor, which is deficient in pernicious anemia, or with an extragastric intrinsic factor, which accounts for the formation of the hemopoietic factor found in the serum of patients with pernicious anemia who have been treated by injections of vitamin B_{12}.[2]

Vitamin B_{12} has been found in liver, muscle, milk, etc., but in order to obtain large amounts it is usually extracted from culture mediums fermented by *Streptomyces griseus*. Apparently this vitamin plays a part in the synthesis of proteins, nucleoproteins, and thymidine, especially in nuclear metabolism (see Chap. 49, Vitamins).

Vitamin B_{12} is usually far more potent than folic acid, but a few cases of megaloblastic anemia (acute macrocytic anemia of children, etc.) respond to treatment with folic acid, but not to treatment with vitamin B_{12}. Pernicious anemia following gastrectomy usually responds less to treatment with vitamin B_{12} than to that with folic acid.

Other nutritive and endocrine factors. Erythropoiesis and the formation of hemoglobin are conditioned by many nutritive factors. Deficiency of several vitamins (ascorbic acid, niacin, riboflavin, pyridoxin, etc.) provoked experimentally produces anemia. Anemia is also observed in patients suffering from scurvy, pellagra, or ariboflavinosis.

There is a more or less severe decrease in erythropoiesis in several endocrine disturbances such as hypophyseal, thyroid, and adrenal insufficiencies. Testicular hormones stimulate erythropoiesis, while ovarian hormones have a slight inhibitory effect. Polycythemia of unknown origin is observed in some cases of hyperfunction of the adrenal cortex and in experimental lesions of the diencephalon.

[1] BERK, L., W. B. CASTLE, *et al.*, *New England J. Med.*, **239**, 111, 1948; *Proc. Staff Meet., Mayo Clin.*, **25**, 105, 1950.

[2] CALLENDER, S. T., and L. G. LAJTHA, *Blood*, **6**, 1234, 1951.

THE LIFE SPAN OF THE ERYTHROCYTE

The average life of the erythrocyte has been estimated in several ways. Erythrocytes from a donor of a different, but compatible, group are transfused into a recipient, and the survival of these foreign erythrocytes is determined by repeated isoagglutination tests. The same procedure can be followed using erythrocytes with radioactive iron in their hemoglobin. These injected cells live from 25 to 150 days; most of the observers who have published their results recently agree that the average life of the erythrocyte is 120 days.

The destruction of the erythrocyte. Red cells are being continuously destroyed. Probably they split up into several fragments which are picked up by the cells of the reticuloendothelial system. The proteins and iron of the erythrocyte are stored and used for the formation of new erythrocytes. A great part of the heme in hemoglobin is transformed into bilirubin, about 40 mg. in each gram of hemoglobin (Whipple). For a daily excretion in the bile of 500 mg. of bilirubin, 12.5 gm. of hemoglobin would have to be destroyed; this corresponds to the hemoglobin in 80 cc. of blood. According to other estimates, the amount of hemoglobin destroyed daily varies from 5.5 to 25 gm. The amount of erythrocytes destroyed can be estimated by (*a*) the quantity of bilirubin excreted in the bile and to some extent by the concentration of bilirubin in the blood or its excretion in the urine; (*b*) by the quantity of stercobilin excreted in the feces.

In cases of marked destruction of hemoglobin, a pigment called hemosiderin accumulates in the liver, the spleen, and other organs. It contains up to 17 per cent iron and gives the prussian blue reaction.

FUNCTIONS OF THE SPLEEN

Up to quite recently the spleen was considered to be an enigmatic organ, but now several of its functions are known. It is not indispensable to life, as it can be removed without causing any important disturbance; its functions are either dispensable or carried out by other tissues.

The following structural features are important: (*a*) it has lymph nodes (malpighian bodies) around the arterioles; (*b*) its veins are connected with venous sinuses and clefts in the splenic pulp; (*c*) there are many phagocytic

cells belonging to the reticuloendothelial system in the splenic pulp; (*d*) the blood comes into intimate contact with the cells of the red pulp.

The spleen as a store or reservoir of blood. In normal conditions the spleen is mainly a reservoir of erythrocytes. The arteries open into the sinuses and clefts in the splenic pulp, which is like a sponge in the meshes of which the erythrocytes are stored and from which they can be ejected later into the venous sinuses. The dilatation and contraction of the spleen is due to its plain-muscle fibers, which are stimulated and inhibited by nerve impulses and by humoral factors.

In some animals the relaxed spleen can contain from 15 to 30 per cent of the blood volume. In the cat and the dog the spleen can contract and relax considerably. In man the spleen is not so important a reservoir as in other animals. To estimate the capacity of the spleen as a blood reservoir, the concentration and total amount of erythrocytes in the circulating blood are measured before and after provoking splenic contraction. Contraction of the spleen has been studied in dogs by the following methods:

1. Small metal bands are clipped on the edges of the spleen, and after the animal has recovered from the operation, changes in the size of the spleen can be followed by taking a series of x-ray pictures.
2. The spleen is placed in a plethysmograph which is left permanently in the abdomen.
3. The spleen is "exteriorized," *i.e.*, taken out aseptically from the abdomen and fixed to the body wall, keeping it protected by a suitable bandage, which is removed in order to observe and measure the changes in its size.

These methods allow the survival of the animals in good condition indefinitely and the observation of the size of the spleen in many physiologic conditions without the use of anesthesia.

Several factors produce rapid contraction of the spleen: (*a*) anoxia due to lack of oxygen, asphyxia, anesthesia, high altitude, or CO intoxication; (*b*) hemorrhage; (*c*) muscular exercise; (*d*) cold; (*e*) emotions and reflexes; (*f*) adrenaline; (*g*) estrus and pregnancy. This contraction is caused mainly by nerve impulses, but humoral factors can also induce it; the denervated spleen contracts when there is a sudden discharge of adrenaline from the adrenal gland, or of sympathin into the blood.

The sudden emission of the erythrocytes deposited in the spleen produces a rapid but transitory increase in the concentration and total volume of erythrocytes in the circulating blood. In splenectomized subjects the increase in erythrocytes provoked by emotion, muscular exercise, or the injection of adrenaline is not so great as in normal subjects; such increase as is observed in these cases is due to the evacuation of other reservoirs or to the loss of plasmatic fluid from the blood.

Hemopoietic function. During fetal life erythrocytes and leukocytes are formed in the spleen. In the normal adult erythrocytes are not produced in the spleen, but this organ has a slight effect on erythropoiesis, which can become very marked in certain pathologic states. Lymphocytes are formed in the malpighian bodies of the spleen, and monocytes in the reticuloendothelial system, but granulocytes are not produced.

In certain pathologic states, foci of myeloid tissue appear in the spleen. In these foci, erythrocytes and granulocytes are formed.

In cases of purpura hemorrhagica with a diminished platelet count (thrombocytopenia), the following signs are observed: (*a*) hemorrhages in the skin and mucosae, and fragility of the capillaries; (*b*) lengthening of the bleeding time; (*c*) decrease in the platelet count; (*d*) failure of the clot to retract after coagulation; (*e*) decrease in granulocytes (neutropenia). Splenectomy causes an immediate increase in the platelets, and capillary fragility disappears. Many patients are greatly improved by this operation. Two interpretations of these facts have been given: (*a*) there is excessive destruction of the platelets by the spleen; (*b*) the spleen secretes an excess of a humoral factor (hypersplenia) which inhibits the output of platelets into the circulation. There is an increase of megakaryocytes (the platelet-producing cells) in the hemopoietic tissues, which is considered as a compensatory reaction by those who sustain the first of these theories, and as the result of an arrested maturation and liberation of the platelets by those who admit the second.[1]

Phagocytic function. The spleen forms part of the reticuloendothelial system. The large cells of the splenic pulp, called histiocytes or macrophagocytes, ingest cells, parasites, worn-

[1] Dameshek, W., and E. B. Miller, *Blood*, 1, 27, 1946.

out erythrocytes, and substances in colloidal dispersion such as hemoglobin. After splenectomy the macrophagocytes in the lymph nodes, liver, and bone marrow become hypertrophied, as well as other lymphoid structures, and splenoid tissue can appear in the lymph nodes and liver.

Destruction of erythrocytes and production of bilirubin. There is no proof that normal erythrocytes undergo phagocytosis in the spleen; but worn-out erythrocytes, those damaged by toxins, hemolysins, or infections, and especially the spherocytes of hemolytic jaundice are captured and destroyed by the splenic phagocytes. After having been deposited for some time in the splenic pulp, some of the erythrocytes have an increased fragility in hypotonic solutions (Banti). Moreover splenectomy diminishes the fragility of the erythrocytes. Botazzi has called this function of capturing and destroying fragile erythrocytes the "hemocatheretic" function of the spleen.

In congenital hemolytic jaundice splenectomy diminishes anemia, jaundice, and the fragility of the red cells. This effect is attributed to the removal of an organ that destroys the erythrocytes, especially the spherocytes.

The spleen, because of its erythrocyte-destroying function, is one of the sites where extrahepatic bilirubin is formed (Mann); this is sometimes called the "biligenic" function of the spleen.

Storage functions. The spleen can store cells, parasites, bacteria, and several chemical substances. When there is an excessive destruction of erythrocytes and hemoglobin, a dark brown pigment, called hemosiderin, is formed; it has a large proportion of iron and gives the prussian blue reaction with potassium ferrocyanide in chlorhydric medium. The normal spleen is one of the stores for iron; ferritin, a crystallized protein with 20 per cent iron, has been obtained from splenic tissue. This store is of special importance for sucklings, as milk contains very little iron.

Abnormal storage (thesaurosis) is observed in certain pathologic conditions. In Gaucher's disease the splenic cells store cerebrosides, and therefore great quantities of kerasine. In Niemann-Pick's disease these cells have a foamlike aspect, due to stored phosphatides. Storage of parasites is observed in malaria and leishmaniasis. Ascoli injected adrenaline in chronic cases of malaria so as to produce contraction of the spleen and the expulsion of the parasites into the circulation, where they could be destroyed by adequate doses of quinine.

Functions in infection and immunity. The spleen increases in size (splenomegaly) in many acute and chronic infections. It takes part in the production of antibodies, as is shown by the fact that some of these diminish after splenectomy. In some cases splenectomy transforms a latent or slight infection into a severe and frequently fatal process. A typical case is the infection of white rats by *Bartonella*. The animals show no signs of infection until the spleen is removed, when a severe infectious condition appears, with intense anemia and a high mortality. If the splenectomized infected rat is joined to a normal rat in parabiosis,[1] the disease is controlled; the spleen therefore acts through a humoral mechanism (Flaum and Lauda).

Splenectomy. Removal of the spleen usually produces anemia in man and in several other species. The anemia is not very severe, and recovery frequently occurs after a few months. In the dog the concentration of erythrocytes decreases by 1,000,000 to 3,000,000 per cubic millimeter; anemia is preceded or followed by reticulocytosis and in a few cases by polycythemia. In the course of this anemia the resistance of the erythrocytes to hemolysis by hypotonic salt solution is increased, and sometimes less iron is retained than normally. In the erythrocytes abnormal structures known as "Jolly's corpuscles" are observed not only after splenectomy but also in cases of atrophy of the spleen. A leukocytosis of 20,000 to 40,000 per cubic millimeter, due to an increase in the granulocytes, is generally present for a short time after splenectomy. There is an immediate increase, of short duration, in the thrombocyte count.

There is no satisfactory explanation of these effects of splenectomy. They seem to be due to a regulatory action of the spleen on the bone marrow through the blood.

[1] Parabiosis consists in establishing a permanent connection between the circulation of two animals of the same species. Incisions, opening the peritoneal cavities, are made in the flank of each animal, and the animals are then sewn together. Vascular anastomoses between the animals are rapidly formed, and there is circulatory exchange between them.

BIBLIOGRAPHY

Hemopoiesis

CASTLE, W. B., and G. R. MINOT, "Pathological Physiology and Clinical Description of the Anemias," Oxford, New York, 1936.

HAM, T. A., W. B. CASTLE, F. H. GARDNER, and G. A. DALAND, Diagnostic Problems in Polycythemia and Anemia, *New England J. Med.*, **234**, 815, 1950.

HURTADO, A., C. MERINO, and E. DELGADO, Influence of Anorexia on Hemopoietic Activity, *Arch. Int. Med.*, **75**, 284, 1945.

JUKES, T. H., and E. L. R. STOKSTAD, Folic Acid, *Physiol. Rev.*, **28**, 51, 1948.

MERINO, C., *Blood*, **5**, 1, 1950 (blood and bile pigments at high altitude).

OSGOOD, E. E., and A. J. SEAMAN, The Cellular Composition of Normal Bone Marrow as Obtained by Sternal Puncture, *Physiol. Rev.*, **24**, 46, 1944.

PONDER, E., W. B. CASTLE, *et al.*, Some Aspects of Red Cell Production, *Ann. New York Acad. Sc.*, **48**, 577, 1947.

RICHES, E. L., *et al.*, *Science*, **107**, 396; **108**, 634, 1948 (vitamin B_{12}).

SHEMIN, D., and D. RITTENBERG, The Life Span of the Human Red Blood Cell, *J. Biol. Chem.*, **166**, 627, 1946.

SMITH, E. L., and C. C. UNGLEY, *Brit. M. J.*, **2**, 1367, 1949; **1**, 151, 1950 (vitamin B_{12}).

SPIES, T. D., *et al.*, *Lancet*, **254**, 239, 1948 (folic acid).

SPIES, T. D., *et al.*, *J. A. M. A.*, **139**, 521, 1949 (vitamin B_{12}).

STOKSTAD, E. L. R., and T. H. JUKES, *Ann. Rev. Biochem.*, **18**, 435, 1949 (folic acid).

WELCH, A. D., and R. W. HEINLE, *Pharmacol. Rev.*, **3**, 345, 1951 (folic acid, vitamin B_{12}).

WHIPPLE, G. H., and S. C. MADDEN, Hemoglobin, Plasma Protein and Cell Protein, *Medicine*, **23**, 215, 1944.

ZUCKER, T. F., and L. M. ZUCKER, *Vitamins & Hormones*, **8**, 1, 1950 (vitamin B_{12}).

Functions of the Spleen

BARCROFT, J., *J. Physiol.*, **60**, 443, 1925.

BARCROFT, J., Die Stellung der Milz in Kreislaufsystem, *Ergebn. d. Physiol.*, **25**, 818, 1926.

KRUMBHAAR, E. B., Functions of the Spleen, *Am. J. M. Sc.*, **184**, 215, 1932.

LAUDA, E., "Physiologie der Milz," Urban & Schwarzenberg, Berlin, 1933.

PERLA, D., and J. MARMORSTON, "The Spleen and Resistance," Baillière, London, 1935.

CHAPTER 6

The Leukocytes and the Platelets

THE LEUKOCYTES

The white cells or leukocytes are found in the blood and the lymph, and in small numbers in the tissues and tissue fluids. They are divided into granulocytes, lymphocytes, and monocytes. The granulocytes or polymorphonuclear leukocytes are classified according to the staining affinity of their granules into neutrophil,[1] eosinophil, and basophil.

Concentration. There are 5,000 to 10,000 leukocytes per cu. mm. in normal blood. When the concentration is persistently less than 5,000 per cu. mm. it is said that there is leukopenia. When the figure remains constant at 10,000 or more, there is leukocytosis; but it must be noted that 11 per cent of normal persons have a concentration slightly above this maximum. The average figure in basal conditions (fasting, and physically and mentally at rest) is 7,000 per cu. mm. The formation and destruction of leukocytes are carried on continuously, and the concentration of leukocytes in blood is the result of an equilibrium between these processes that is a true homeostatic mechanism.

The concentration of leukocytes is not the same in the large blood vessels as in the capillaries. When the circulation is slowed down, the number of leukocytes in the small vessels increases; the cells go toward the periphery of the blood stream, near the endothelium (margination of the leukocytes). In the state of shock provoked by the injection of peptone or histamine, and in anaphylactic shock, there is an abnormal distribution of the leukocytes; there is leukopenia in the peripheral blood and leukocytosis in the blood in the abdomen.

[1] In some mammals, amphophil (staining with both acid and basic dyes) instead of neutrophil granules are found.

Table 7. Concentration of Leukocytes in the Blood

Leukocytes	*Absolute concentration of leukocytes, number per cu. mm.*		*Relative concentration of leukocytes, %*	
	Range	*Average*	*Range*	*Average*
Granulocytes				
Neutrophil. .	3,000–7,000	4,300	55–65	65
Eosinophil. .	50–500	200	1–4	2
Basophil. . . .	0–50	25	0–1	0.5
Lymphocytes. .	1,000–3,000	2,000	20–30	27.5
Monocytes. . . .	100–600	400	4–10	5
Total.	5,000–10,000	7,000	. . . . .	100

The percentage of the different types of leukocytes is called the relative count.

The increase in concentration of only one type of cell is known, according to the cells involved, as neutrophilia, eosinophilia, basophilia, lymphocytosis, or monocytosis. A decrease in concentration is known as neutropenia, lymphopenia, or monocytopenia. The variations in the concentration of the different leukocytes cannot be appreciated by examining only the relative count or percentage of cells; only the absolute count can give an accurate picture. Thus in a case of leukocytosis with 15,000 leukocytes per cu. mm., where the percentage of lymphocytes is only 15, there is no lymphopenia, because the total number of these cells per cu. mm. is 2,250, which is a normal figure.

In newborn infants a high count of 20,000 or more leukocytes per cu. mm. is usually ob-

served, but the concentration rapidly falls to a normal figure. In children lymphocytes are more numerous (50 per cent) than in adults; they diminish gradually so that the adult figures are reached by puberty (Fig. 15). In pregnancy there is slight neutrophil leukocytosis. Violent exercise, emotion, pain, and the injection of adrenaline provoke a rapid increase in the leukocyte count, due mainly to the evacuation of leukocytes from the blood stores. In the course of a single day considerable fluctuations in the white cell count have been found in the same subject; in some cases variations of 2,000 per cu. mm. in 1 hr. have been observed. The so-called digestive leukocytosis seen after the midday meal has been recently considered not as caused by the ingestion of food but simply as a slight rhythmic increase (of about 2,000 per cu. mm.) occurring in the afternoon, while in the morning lower figures are usually found.

FIG. 15. Neutrophil and lymphocyte count at different ages. Shaded areas correspond to the range of variation in normal adults.

Neutrophilia, causing a leukocytosis of 20,000 to 40,000 and more, is frequently observed. It is found (*a*) in infection by bacteria, especially by cocci, fungi, viruses, and parasites; (*b*) in cases where there is suppuration, such as abscesses or appendicitis; (*c*) in many intoxications; (*d*) after injection of foreign proteins; (*e*) transitorily after hemorrhage.

Eosinophilia is observed in (*a*) infestation by parasites, such as trichinosis, hydatidosis, hookworms, etc.; (*b*) allergic states, such as asthma; (*c*) certain cutaneous diseases, etc.

Leukopenia is observed in (*a*) anaphylactic shock and shock induced by the injection of peptone, both conditions in which the leukocytes are accumulated in the small vessels of the abdomen;[1] (*b*) certain infectious conditions produced by bacteria (typhoid, sepsis, and tuberculosis), viruses (influenza), and protozoa (malaria); (*c*) cases in which hemopoiesis decreases (aplasia of the bone marrow, etc.); (*d*) certain intoxications (benzene, pyramidon, sulfonamides, thiouracil, etc.); (*e*) exposure to radioactivity. In some cases the leukocytes diminish considerably, the white cell count remains consistently below 2,000 or even less, and there is severe infection of the throat; many of these cases end fatally. This condition is known as agranulocytosis. Injections of pentose nucleotide have been used to stimulate granulopoiesis and of penicillin to prevent or cure serious infections that may occur in these cases. Eosinopenia has been observed after the injection of the adrenocorticotrophic hormone of the hypophysis in subjects with intact adrenals, the maximum decrease occurring 4 hr. after the injection. In patients in whom the adrenal glands have been damaged (Addison's disease) this response was not observed; therefore it is due to an increase in the corticoadrenal secretion. Cortisone, one of the steroids of the adrenal cortex, also provokes eosinopenia.[2]

Lymphocytosis occurs in a few acute infectious diseases (*e.g.*, whooping cough); it is more common in chronic infections (*e.g.*, syphilis). In the acute stages of tuberculosis there is lymphopenia, the decrease in leukocytes being greater in the more severe forms; when the patient improves, lymphocytes increase up to the normal figure or a little above. Lymphopenia is also one of the signs of Selye's "alarm reaction" to trauma or sudden stress. The adrenocorticotrophic hormone of the hypophysis seems to play a part in regulating the concentration of lymphocytes in the blood. When the secretion of this hormone increases, it provokes hypersecretion of the corticoadrenal hormones, which produces atrophy of the lymphoid tissues and destruction of some of the cells; there is a marked decrease in the circulating lymphocytes and eosinophils.

Lymphopenia seems to increase the severity

[1] After a few hours leukopenia is followed by leukocytosis.

[2] FORSHAM, P. H., *et al.*, *J. Clin. Endocrinol.*, 8, 15, 1948.

of infections and favor the development of experimental cancer (Murphy). These facts suggest that the lymphocytes play an important part in immunity.

The origin of the leukocytes. The granulocytes are formed in the bone marrow; they are therefore of myeloid origin (Table 5). A primitive cell, the myeloblast, develops to a promyelocyte, which gives rise to the myelocytes, cells with a rounded nucleus and specifically staining granules. The nucleus indents, takes the shape of a crook (metamyelocyte), and then becomes lobed (granulocyte). The lymphocytes are formed in the germinal centers of the lymph nodes and other lymphoid tissues. The monocytes have their origin in the histiocytes of the reticuloendothelial system, principally in the liver, and to a lesser degree in other hemopoietic organs.

Neutrophil leukopoietic activity can be estimated by examining the aspect of the nuclei of the neutrophil granulocytes in the blood and establishing the percentage of each type. Arneth's classification (1904), or more usually Schilling's simpler one, is used. Young, recently formed granulocytes have an indented nucleus; as they grow older, lobulations appear and increase in number. There is said to be a shift to the left when young forms with nuclei in the shapes of a U, V, or L predominate, and a shift to the right when the older cells with lobulated nuclei are more numerous (Table 9).

Table 8. Blood Picture of Several Species (Average Values)

Species	*Er*	*Hb*	*PCV*	*MCV*	*MCHb*	*MCHbC*	*MED*	*L*	*N*	*E*	*B*	*Ly*	*M*
Human													
Man	5.4	15.4	47	87	29	34	7.5	7	65	2	0.5	27.5	5
Woman	4.8	13.8	42	87	29	34	7.5	7	65	2	0.5	27.5	5
Horse	6–8	11–15	35	54	18	34	6.0	7.5	56	4	0.5	31.5	8
Ox	6.3	12	40	58	20	34	5.8	8.0	25	5	0.5	63.5	5
Sheep	10.5	12.5	37	35	12	32	4.7	16.5	30	8	0	60	2
Goat	14	10	32	18	8	32	4.1	9	40	2	0	54	2
Dog	6.5	13	39	59	20	32	7.2	11	71	5	0.5	21.3	2
Cat	7.8	11	40	57	15	29	6.5	13	59	5	0.5	32.5	2
Rabbit	6.2	13	39	64	21	33	6.7	7.8	52	3	2	34	5
Guinea pig	5.8	14	48	83	25	31	7.1	8	42	4	0.5	45.5	8
Rat	6.8	13	48	61	20	33	6.3	9	20	2	0.5	73	5

Er = erythrocytes, millions per cu. mm.
Hb = hemoglobin, gm. per 100 cc. of blood.
PCV = volume of packed cells, cc. per 100 cc. of blood.
MCV = mean corpuscular volume, cu. μ.
MCHb = mean corpuscular hemoglobin, μg.
MCHbC = mean corpuscular hemoglobin concentration, gm. of hemoglobin in 100 cc. of erythrocytes.
MED = mean erythrocyte diameter, μ.
L = leukocytes, thousands per cu. mm.
N = neutrophil granulocytes, per cent.
E = eosinophil granulocytes, per cent.
B = basophil granulocytes, per cent.
Ly = lymphocytes, per cent.
M = monocytes, per cent.

Motility. The granulocytes have ameboid movements; they advance by sending out pseudopodia. The other leukocytes are less motile. Nevertheless crawling movements have been described in the lymphocytes, and the

Table 9. Schilling's Index. Percentage of Different Types of Leukocytes for Normal Blood

Neutrophils							
Myelocytes	*Young metamyelocytes**	*Adult myelocytes†*	*Lobed nucleus‡*	*Eosinophils*	*Basophils*	*Lymphocytes*	*Monocytes*
0	0–1	3–5	55–65	2–4	0–1	20–25	4–8

* Indented nucleus.
† U-shaped nucleus.
‡ Two, three, four, or five well-marked lobulations.

monocytes sometimes show an undulatory movement of their contour (Carrel).

Tactisms. Certain substances attract (positive chemiotaxis) and others repel (negative chemiotaxis) the leukocytes, as can be demonstrated by placing under the skin small capillary tubes containing the substances. Several hours after broth has been injected into the peritoneum or the pleura, fluid with abundant leukocytes can be removed by puncturing the cavity.

Diapedesis. Microscopic examination of living tissues extended on a warm plate (37°C.) shows the leukocytes migrating from the blood vessels after previous margination; they pass out from the capillaries through the interstices between the endothelial cells. This is known as diapedesis.

Enzymes. Neutrophil granulocytes contain several enzymes. There is a protease which acts in an alkaline or neutral medium; it is inhibited by an antienzyme in the plasma. When large quantities of granulocytes are accumulated and disintegrated, the enzyme thus set free digests the tissues (abscesses). Oxidases are found in myeloid cells; at one time they were considered specific for this type of cell, but they have also been found in other nonmyeloid white cells. Lipases, acting on butterfat and on waxes, are found in the lymphocytes. Leukocytes also contain enzymes that digest bacteria (lysozymes). Carrel has given the name of trephones to certain substances (amino acids and peptones) in the leukocytes that stimulate the growth of tissue cultures.

Phagocytosis. Metschnikoff (1883) discovered this important function of the leukocytes, which consists in the ingestion of cells and particles into the protoplasm. The granulocytes, especially the neutrophils, are microphagocytes, *i.e.*, they ingest bacteria and small particles. The monocytes and the histiocytes in the tissues are macrophagocytes, *i.e.*, they ingest larger particles, such as worn-out erythrocytes and cells. These macrophagocytes are mesenchymatic cells, which retain the embryonic capacity of development and have in common the property of phagocytosis. The phagocytes fixed in the tissues form the reticuloendothelial system (Table 10); the term "endothelial" is not quite accurate as these cells are not endothelial cells but mesenchymatic phagocytes.

Table 10. Reticuloendothelial System

Reticuloendothelial cells, or histiocytes

- Fixed histiocytes
  - Connective tissue → Fixed histiocytes
  - Reticulum → Reticulum of the spleen, lymph nodes, bone marrow
  - Endothelium → Sinusoids in liver (Kupffer cells), spleen, bone marrow
  - Microglia
- Circulating histiocytes
  - Tissues → Migrating histiocytes
  - Blood → Monocytes

Phagocytosis is carried out in several steps: (*a*) contact between the particle and the phagocyte; (*b*) ingestion of the particle; (*c*) digestion, if the substance is a digestible protein (*e.g.*, bacteria); if the substance resists the action of the phagocyte, the latter serves as a carrier for the former, or else disintegrates. Several conditions are necessary for phagocytosis to take place: (*a*) contact with the particle must be established; (*b*) there must be an adequate temperature (38°C.), oxygen pressure, and ionic equilibrium (Ca^{++} is necessary). Many factors modify phagocytosis: (*a*) most anesthetics inhibit it; (*b*) thyroid hormone stimulates it, and it diminishes in thyroid insufficiency (thyroidectomy or hypophysectomy); (*c*) it is partially or totally inhibited by certain bacterial products called "aggressins"; (*d*) it diminishes when there is a deficiency of certain vitamins, ascorbic acid, thiamine, or pyridoxin.

Blood serum contains substances which, by acting on bacteria, favor phagocytosis (Denys, 1895); they were called "opsonins" by Wright (1903). Immunized animals have specific antibodies, called "bacteriotropins," which specifically favor phagocytic activity against the bacteria to which the animal has been immunized; these antibodies resist heating and need the presence of the complement (see Chap. 9).

Significance of phagocytosis. The particles and substances taken up by the phagocytes are submitted to a process of intracellular digestion, carried out by means of the enzymes in the cell. Indigestible particles (charcoal) and certain highly virulent bacteria resist this process. Some of the invertebrates normally digest food by phagocytosis; the cells of their digestive tract capture alimentary particles and digest them. Phagocytosis also takes part in the reabsorption of embryonic tissues; *e.g.*, the tail of tadpoles. Finally, phagocytosis plays a fundamental part in immunity—the capture and digestion of germs, a function carried out mainly by the neutrophils (microphagocytes) and histiocytes (macrophagocytes).

Menkin[1] describes several factors of importance in inflammation: (*a*) leukotaxine, a crystalline substance, probably a polypeptide, which increases capillary permeability and the outward migration of leukocytes; (*b*) a factor that stimulates leukocytosis (LPF), a pseudoglobulin (α_2 globulin) destroyed by heat, produced in inflamed tissues, which stimulates the production of granulocytes in the bone marrow; (*c*) necrosin, a toxic euglobulin, which destroys the cells of inflamed tissues, and causes lesions in the liver and kidney if it passes into the circulation; (*d*) pyrexin, a substance which produces fever.

THE PLATELETS OR THROMBOCYTES

These constituents of the blood were described by Bizzozero (1880), who called them "platelets"; they are also known as "thrombocytes." Other names for them are no longer in use ("Hayem's hematoblasts," "globulins," etc.). They exist in the circulating blood and are not artificially formed *in vitro*, as was at one time believed.

Properties of the platelets. They are extremely fragile and have a tendency to stick on any surface (slides, etc.) and to agglutinate together; they are easily deformed and destroyed. Anticoagulants (sodium citrate, heparin) added to the blood help to delay the destruction of the platelets *in vitro*. When kept in good condition they are oval or spindle-shaped, without a nucleus, and from 2 to 4 μ in diameter. They have a low specific gravity and float in the plasma on sedimentation of the erythrocytes, but they can be separated by prolonged centrifugation. The dried platelets have 60 per cent protein and 15 per cent fat; in the ash there is P, Fe, K, S, and Ca. They reduce methylene blue to the leukobase, and they have a measurable oxygen consumption, but their metabolism is not well known. Apparently they have a life span of only a few days.

Origin and destruction. The platelets are formed by the megakaryocytes of the bone marrow. The spleen plays a part in the destruction and possibly in the liberation of these cells (see "Functions of the spleen," Chap. 5).

Concentration. There are about 250,000 (150,000 to 400,000) platelets per cu. mm.; by some methods of counting, from 600,000 to 900,000 are found, but these high figures are probably due to artificial fragmentation. Thrombopenia, or decrease in the platelet count, occurs in (*a*) acute infections; (*b*) peptone and anaphylactic shock, in which there is at first a fall in their number (destruction or redistribution) but soon afterward they reappear; (*c*) certain hemorrhagic diseases, *e.g.*, thrombocytopenic purpuras; (*d*) aplastic anemia; (*e*) relapse of pernicious anemia.

Functions. Platelets easily adhere to particles (India ink, bacteria) floating in the plasma, both *in vitro* and in the circulating blood, a process that seems to favor phagocytosis. Platelets are an important factor in blood clotting: (*a*) they form the knots in the fibrin reticulum; (*b*) they release substances that activate the cofactor for thromboplastin and convert thromboplastinogen into thromboplastin (see Chap. 8); (*c*) on adding platelets to blood, the clot retracts markedly, and it fails to retract when antiplatelet serum is given (Le Sourd and Pagniez); (*d*) there is thrombopenia in several, but not in all (Roskam), of the hemorrhagic diseases with a disturbance in blood coagulation (see Chap. 8, Coagulation of the Blood), *e.g.*, in thrombocytopenic purpura.

When a tissue is wounded the platelets agglutinate and tend to obstruct the small damaged blood vessels. They set free thromboplastin, a blood-clotting factor, and vasoconstrictor substances, which act on the smooth-muscle fibers.

Splenectomy is followed by a transitory increase in the platelet count in healthy animals. In cases of thrombocytopenic purpura there is a

[1] MENKIN, V., *Arch. Path.*, **41**, 376, 1946.

sustained increase in the platelet count after splenectomy, and the hemorrhagic symptoms improve; the clot also becomes firmer.[1]

Hyperplasia of megakaryocytes in the bone marrow has been observed in cases of thrombocytopenic purpura. This has been attributed to (*a*) a compensatory reaction to excess phagocytosis of thrombocytes by the spleen; (*b*) inhibition of maturation and release of thrombocytes owing to an increase in the inhibitory functions of the spleen, *i.e.*, hypersplenia (Dameshek and Miller, see "Functions of the spleen" in Chap. 5).

BIBLIOGRAPHY

Leukocytes and Phagocytosis[2]

BUNTING, G. H., *Physiol. Rev.*, **2**, 505, 1922.

BUNTING, G. H., in H. Downey, "Handbook of Hematology," vol. 1, Hoeber, New York, 1938, p. 437, (functions of the leukocytes).

FLEISCHMANN, W., *Ergebn. d. Physiol.*, **27**, 1, 1928 (properties of leukocytes).

[1] This improvement is not observed when there is bone-marrow aplasia; this should be determined by sternal puncture before the operation.

[2] For textbooks of hematology, see Bibliography, Chap. 1.

GARREY, W. E., and W. R. BRYAN, *Physiol. Rev.*, **15**, 597, 1935 (variations in leukocytes).

HAMBURGER, F., in Abderhalden's "Handbuch für biologischen Arbeitsmethoden," Abt. IV, **54**, 952, Urban & Schwarzenberg, Berlin, 1925 (phagocytosis).

LORENZ, E., *Anal. Inst. Clin. Med.*, **2**, 181 1917 (differential count).

MCCUTCHEON, M., *Physiol. Rev.*, **26**, 319, 1946.

MUDD, C., M. MCCUTCHEON, and B. LUCKE, *Physiol. Rev.*, **14**, 210, 1934 (phagocytosis).

PHILIPSBORN, E., *Klin. Wchnschr.*, No. 9, 373, 1926 (phagocytosis).

Reticuloendothelial System

DU BOIS, A. H. "Physiologie et physiopathologie du système reticuloendothélial," Masson et Cie, Paris, 1934.

JAFFÉ, R. H., *Arch. Path. Lab. Med.*, **4**, 45, 1927.

JAFFÉ, R. H., *Physiol. Rev.*, **11**, 277, 1931.

PITTALUGA, G., "Las Enfermedades del Sistema Retículo-endotelial," Espasa-Calpe, Madrid, 1934.

Thrombocytes

AYNAUD, M., Les Globulins, in Gilbert and Weinberg, "Traité du sang," vol. 1, J. C. Baillière et fils, Paris, 1913, p. 410.

ROSKAM, J., "Physiologie normale et pathologique du globulin," Presses Universitaires de France, Paris, 1927.

TOCANTINS, L. M., *Arch. Path.*, **23**, 850, 1937.

TOCANTINS, L. M., *Medicine*, **17**, 155, 1938.

CHAPTER 7

Hemorrhage and Transfusion. Blood Groups

HEMORRHAGE

Hemorrhage is the escape of blood from the blood vessels. The blood is lost either outside the body (external hemorrhage), or into the tissues and body cavities (internal hemorrhage). The hemorrhage can be arterial, venous, or capillary, according to the nature of the damaged vessel; it can be produced suddenly or slowly and can be moderate or severe according to the amount of blood lost. It can be due to natural causes or can be artificially produced. Hemorrhage due to natural causes can be divided into (*a*) traumatic (road, industrial, and other accidents, war, etc.); (*b*) medical (gastric and duodenal ulcers, pulmonary tuberculosis, etc.); (*c*) surgical (operations, etc.); (*d*) obstetric (childbirth, miscarriage, etc.). Hemorrhage is provoked (*a*) for experimental purposes; (*b*) to obtain blood for transfusion; (*c*) for therapeutic purposes.

Amount of blood that can be lost. Experimental observations on laboratory animals are more numerous than accurate observations in man, but recently these also have been made in volunteers. Sudden hemorrhage should be distinguished from slow or repeated hemorrhage; the former is more severe than the latter. Small hemorrhages can be repeated many times, so that in the course of months or years great quantities of blood are lost; in the intervals between the bleedings, the blood is regenerated. A sudden loss of blood equivalent to 1 per cent of the body weight is well tolerated by animals of many species. A loss equivalent to 3 per cent of the body weight causes death in a large percentage of rabbits and a small percentage of dogs. The majority of dogs die when they lose 55 per cent of the blood volume (Richet). In man, loss of 30 to 40 per cent of the total blood volume endangers life and should be treated without delay by blood transfusion.

Signs and symptoms of hemorrhage. Signs and symptoms of hemorrhage appear and increase in severity as the amount of blood lost increases. At first there is only a slight fall in blood pressure and a more rapid pulse if the patient is standing; if he is lying down there may be no signs or symptoms. A greater loss of blood will cause disturbances even if the patient is lying down. The blood pressure falls, the skin and mucosae are pale, and the hands are cold and moist. Later, generalized sweating may be observed, the pulse slows down, and breathing becomes fast and shallow. The patient is restless and feels cold, especially his hands and feet; there is general malaise and a sensation of weakness and dizziness with nausea, which sometimes leads to vomiting. If a very large amount of blood has been lost, the blood pressure is very low, there is a small, threadlike pulse, and respiration is considerably disturbed (air hunger). The patient responds slowly to stimulation and his eyesight is impaired. In the final stages consciousness is lost, the pupils are dilated, the sphincters relax, convulsions may occur, and the patient dies. Even a small loss of blood may provoke alarming symptoms because of emotional stress: paleness, rapid pulse, disturbed breathing with increased pulmonary ventilation, dizziness, and sometimes fainting. Bloodletting should, therefore, always be performed with the patient lying down, not standing up or even sitting.

Children and old people are more sensitive to

hemorrhage than adults. Cold, trauma, and general weakness diminish the resistance to hemorrhage. Even the loss of a small amount of blood is dangerous in shock or adrenal insufficiency.

If 500 cc. of blood is rapidly extracted from a vein, the systolic blood pressure falls only slightly and there is a 15 to 30 per cent increase in the pulse rate; the cardiac output does not vary. If the amount removed is increased to 1 liter, the blood pressure falls and there is usually a slow pulse when the subject is lying down; if he rises the pulse rate increases, the pressure falls even more, and he may faint. Sometimes fainting is due to emotional stress (fear), but the loss of blood alone may be sufficient cause. The frequency of fainting increases as the amount of blood lost is greater. Thus in a series of subjects who lost about 540 cc. of blood, fainting was observed in 8.5 per cent; in another series in which the blood lost was from 800 to 1,000 cc., 39 per cent fainted, and 52 per cent of a series which lost 1,000 to 1,200 cc.

Rapid hemorrhage in anesthetized animals causes a fall in blood pressure and always increases the pulse rate. In nonanesthetized human subjects these signs are not necessarily present in cases of severe hemorrhage. The blood pressure and cardiac output may remain normal for a time, thanks to compensatory reactions; moreover, bradycardia is more frequently observed than tachycardia. The blood pressure is maintained at a high level as a result of the increase in the peripheral resistance produced by constriction of the arterioles. This vasoconstriction is also the cause of paleness and the sensation of cold. When there is a sudden fall in the peripheral resistance, brought about by dilatation of the arterioles, the blood pressure diminishes and the circulation of the brain is deficient; this causes dizziness or fainting.

Effects of hemorrhage. The decrease in the blood volume diminishes the circulation in the tissues, and the oxygen supply becomes insufficient (anoxia). This lack of oxygen is first felt by the heart and the central nervous system; when anoxia is prolonged, other organs (liver, kidney, etc.) also suffer damage and their functions are disturbed. Prompt replacement of the blood lost by transfusion produces a spectacular recovery, even when the condition of the patient is extremely serious; but if the blood pressure remains low for a certain time a state of irreversible shock is established, in which blood transfusion produces only a transitory improvement (sometimes no improvement at all) and the patient dies.

Analysis of the symptoms. In moderate hemorrhage, the arterial blood pressure is maintained by constriction of the arterioles, governed primarily by the vasomotor center and to a much lesser degree by humoral factors (secretion of adrenaline and renin).

The venous return of blood to the heart, and therefore the cardiac output, are not diminished by small hemorrhage (up to 500 cc.). A greater loss causes a decrease in the venous return; therefore the venous pressure and the pressure in the right auricle fall, and both the stroke volume and the minute output diminish. In consequence, the arterial blood pressure falls.

The increase in pulse rate and respiratory frequency is due to reflexes provoked by arterial hypotension. Bradycardia (slow pulse) seems to be caused by a reflex having its origin in the peripheral blood vessels, which causes the center of the vagus to discharge impulses and thus slow down the pulse rate; it is called the vasovagal reflex.

Compensatory mechanisms. The following reactions occur as soon as blood is lost:

1. The blood clots, sealing the damaged blood vessels and thus stopping the hemorrhage. This effect of the clot is favored by a low blood pressure and local vasoconstriction (see "Hemostasis" in Chap. 8).
2. Vasoconstriction in the cutaneous and splanchnic areas maintains the level of the blood pressure, or restores it if the pressure has fallen. This also redistributes the circulating blood, so that a larger amount goes to the central nervous system, heart, and lungs, and less to the skin and mucosae, which are pale and cold.
3. Blood stores discharge erythrocytes into the circulation. The spleen plays an important part as a blood reservoir in the dog; its importance in man is still under discussion.
4. The heart rate increases in some cases (see above), and if there is a sufficient venous return, the cardiac output is also increased.

In cases where there is considerable loss of blood, tissue fluid is reabsorbed into the circula-

tion and dilutes the blood. Hemodilution is directly proportional to the loss of blood and reaches its maximum 10 to 24 hr. (exceptionally, 90 hr.) after hemorrhage. This reabsorption of tissue fluid is due to the osmotic pressure of the plasma proteins, which become more efficient for the attraction of water and salts from the tissues into the capillaries because of the fall in blood pressure. The erythrocyte concentration does not vary immediately after hemorrhage, but begins to diminish when the tissue fluid is reabsorbed into the circulation and continues to fall as hemodilution progresses. Hemodilution can be estimated and its changes followed by repeated erythrocyte counts, or readings of the hematocrit, or determination of the specific gravity of the blood. This reabsorption of tissue fluid is probably the cause of the intense thirst in patients who have suffered a severe hemorrhage.

In a later stage the plasma-protein concentration diminishes (hypoproteinemia) because of the loss of plasma and the dilution caused by reabsorption of water from the tissues. Hypoproteinemia is accompanied by a decrease in the osmotic pressure of the plasma, and therefore the reabsorption of fluid from the tissues decreases; also it causes alterations in the tissue cells and retards healing of wounds. If hypoproteinemia is marked, it should be compensated by transfusion of blood or plasma. During the first day following hemorrhage, fibrinogen and other proteins are rapidly regenerated, but complete recovery takes several weeks (see "The formation of plasma protein" in Chap. 1).

The red and white blood cells are also replaced. A few hours after hemorrhage there is a transitory increase in the leukocyte count (neutrophil granulocytes) to a concentration above normal. The lost erythrocytes are gradually replaced, and a normal concentration is restored in 2 to 3 weeks in healthy subjects.

The physiologic basis of the treatment of hemorrhage will be considered when discussing transfusion.

TRANSFUSION

Transfusion is the injection of whole blood, or some of the constituents of blood, into the circulation. Therapeutic use of transfusion is one of the great achievements of modern treatment, especially in the prevention and cure of surgical and traumatic shock. Its application before, during, and after operations has made possible many procedures which, owing to their severity or long duration, would overcome the resistance of the patients if they were not supported by the effects of transfusion. It should not be left until the condition of the patient has deteriorated to an advanced degree. Early transfusion has no contraindications—indeed, it can be and is used as a preventive measure when there is danger of shock—whereas if transfusion is delayed it may be completely inefficacious.

Transfusion is performed for many and varied purposes. Its main indications are (*a*) *restitution and maintenance of blood volume*, *e.g.*, in hemorrhage, shock, or burns; before, during, and after surgical operations; (*b*) *restitution of plasma volume and plasma-protein concentration*, either because of loss of circulating plasma, as occurs in burns where there is abundant transudation of plasma, or because of deficient production of proteins, as in nutritional hypoproteinemia, dystrophic disturbances in infants, etc.; (*c*) *restitution or maintenance of electrolyte and acid-base equilibriums;* (*d*) *compensation of a decrease in erythrocytes or hemoglobin*, *e.g.*, anemia, carbon monoxide intoxication, methemoglobinemia, etc., or to stimulate hemopoiesis; (*e*) *provision of factors to correct disturbances in blood coagulation;* (*f*) *provision of immune antibodies* for the prevention or treatment of infectious diseases, *e.g.*, transfusion of plasma from convalescent or immune subjects to patients with an active infection; (*g*) *replacement of hemolizable erythrocytes*, *e.g.*, in hemolytic disease of the newborn.

The decrease in blood volume produced by hemorrhage, shock, burns, and other conditions, causes disturbances which may rapidly become fatal. The main disturbances are

1. Deficiency in circulating fluid, which causes a decrease in the venous return to the heart and therefore in the cardiac output. Consequently, the peripheral circulation is slowed down.
2. Deficiency of nutritive and respiratory substances in the tissues, due to retarded circulation caused by hypovolemia and constriction of the arterioles. The deficiency in oxygen (hypoxia or anoxia) thus caused provokes an increase in capillary permeability, which results in the loss of plasma fluid into the tissues. The blood volume diminishes even more, and the circulation becomes still more retarded.

3. Disturbances in the functional activity of the tissues because of oxygen deficiency. Those provoked in the central nervous system and heart are outstanding, but if the condition is prolonged, nearly all the cells and tissues of the organism are disturbed.

Transfusion is performed by injecting into a vein (*a*) whole blood; (*b*) plasma or serum; (*c*) serum albumin; (*d*) erythrocyte suspensions; (*e*) colloid solutions; or (*f*) crystalloid solutions (salts, glucose, amino acids, protein hydrolysates).

Blood can be transfused into the marrow of the sternum (Tocantins). Intra-arterial transfusion is not used in medical practice, although it has some theoretical advantages and in certain experiments it has proved more efficacious than intravenous transfusion in the control of shock.

In order to have blood or other suitable material available in cases of emergency, in which transfusion should be performed as soon as possible, blood stores are established in hospitals and field ambulances with preserved blood or dried plasma. "Blood banks" have ready all types of blood, adequately classified. The supply is renewed continually, thanks to voluntary donors. Blood is usually kept at 4°C. with citrate and glucose added; for example, 500 cc. of blood is mixed with 75 cc. of a fluid containing trisodium citrate $2H_2O$, 1.6 gm.; citric acid H_2O, 0.56 gm.; glucose, 1.5 gm.; and distilled water, 100 ml. After a few days in cold storage the erythrocytes release potassium and gradually disintegrate. Special treatises on blood transfusion should be consulted for the technique of the correct use of different kinds of blood.

Transfusion of Whole Blood[1]

Transfusion of whole blood has the advantage of restoring all the constituents of the circulating fluid, but it has certain drawbacks. There are difficulties in obtaining, conserving, and transporting blood; moreover, the compatibility of the donor's and the receptor's bloods must first be established. Transfusion of blood plasma is often used, especially in wartime, because although blood plasma is not the exact equivalent of whole blood, it fulfills many of its functions and is easier to keep and transport.

[1] Transfusion of whole blood was first performed by Lower (1665) in dogs. It was applied to man by Denis (1667). After Landsteiner's discovery of the blood groups (1901) and L. Agote's application of citrate to prevent blood clotting (1914), transfusion became common in medical and surgical practice.

Transfused blood remains in the receptor's circulation for some time. In some cases 50 per cent of the transfused erythrocytes have been found 45 days after they were injected. Transfused plasma proteins are utilized for the nutrition of the receptor's tissues.

The technique most commonly used consists in the intravenous injection of blood in which clotting is prevented by adding tribasic sodium citrate in a concentration of 0.3 per cent. It is advisable to filter the blood just before using it, in case small clots have been formed. The donor must be lying down quietly. Blood is then drawn from one of the arm veins in the amount of 400 to 500 cc.; not more than one-tenth of the blood volume, or 0.8 per cent of the body weight, should be drawn. The blood should be injected slowly into the recipient, drop by drop, but if the blood volume must be quickly reestablished, as in severe hemorrhage, when the life of the patient is in danger, up to 50 or even 100 cc./min. may be given. The blood pressure should be measured frequently in order to decrease the rate of transfusion as soon as it rises. A too-rapid transfusion may cause dilatation of the heart and diminish its efficiency, especially if the heart is diseased. Transfused blood may stagnate in the capillaries, without increasing the circulating blood volume, as occurs in advanced shock.

Three types of danger must be avoided:

1. Errors in technique (gaseous embolism, etc.) or excessive speed of injection.
2. The transmission of diseases, such as malaria, syphilis, or other infections. The donor should be carefully examined and should not be suffering from allergy.
3. Incompatibility of the donor's and recipient's bloods. If the donor's erythrocytes are agglutinated by the recipient's plasma, serious and even fatal accidents can occur.

A large number of accidents observed immediately after transfusion are due to impurities in the citrate solution, especially to pyretogenous substances in the water used in making these solutions. Accidents have also been attributed to the quality of the glass of the receptacle in which the blood is collected and stored; it should be neutral glass, scrupulously clean and sterilized.

Blood transfusion is absolutely necessary, and should be performed without delay, in cases of hemorrhage in which the loss of blood is from 4 to 5 per cent of the body weight; it is advisable when the loss is 2 to 3 per cent, and is usually dispensable when only 1 per cent or less is lost.

Transfusion of Plasma, Serum, and Serum Albumin

During the Second World War, blood plasma (United States forces) and serum (British forces) were widely used in transfusion as substitutes for whole blood. Serum should not be administered immediately after it has been prepared, because then it is toxic, but it soon loses its toxicity. In cases of hemophilia, however, fresh serum containing thrombin has been given (see Chap. 8). Plasma has many advantages over whole blood except in cases of anemia, in which it cannot be substituted because of the need of erythrocytes, and in disturbances of blood coagulation, in which whole blood is more efficacious. It can substitute for whole blood in cases of hemorrhage and shock, and it is better than blood in cases of burns when there is hemoconcentration; it is also indicated in hypoproteinemia. Fluid plasma can be kept for a long time at 4°C.; at this temperature microorganisms cannot develop. If it is to be transported, it is even better to freeze it at -10 to -20°C. Dried plasma is prepared by freezing and drying in a vacuum (lyophilization) so that it redissolves easily in water. This preparation not only occupies very little space and has a relatively small weight, but it can be kept indefinitely and rediluted at any desired concentration. Plasma has the advantage over whole blood that no blood grouping is needed; moreover, the plasma of about fifty donors is pooled in order to avoid a high concentration of any particular agglutinin.

Serum albumin. Crystallized human serum albumin is less frequently used, but in some cases it has certain advantages. It is the most soluble and stable of the plasma proteins. It is well tolerated in transfusion, better than serum globulins, which in some cases cause accidents. The viscosity of a 25 per cent solution is not greater than that of blood, yet this solution has a strong osmotic pressure which attracts water from the tissues into the blood vessels; 80 per cent of the osmotic pressure of plasma is due to its serum albumin. It has been used with good results in cases of shock and of edema with transudation into the body cavities, *e.g.*, in cases of cirrhosis of the liver and of nephrosis. Since 1 gm. of serum albumin is the equivalent of 18 to 20 cc. of plasma, 25 gm. of serum albumin dissolved in 100 cc. of isotonic sodium chloride solution is the equivalent of 500 cc. of plasma; this amount is known as one unit. Serum albumin solution has the following drawbacks: (*a*) it has no antibodies; (*b*) owing to its high osmotic pressure, it might attract an excess of fluid into the blood vessels and thus overload the circulation.

Serum albumin of cattle and horses is usually well tolerated; occasionally it provokes a reaction in the organism, and it should not be injected more than once because specific antibodies are formed after the first injection (see Chap. 9).

Transfusion of Other Solutions

Fluids used as blood substitutes in transfusion should fulfill the following conditions: (*a*) they should have no toxicity; (*b*) they should form colloid dispersions with an osmotic pressure sufficient to retain fluid within the blood vessels, thus maintaining the blood volume; (*c*) they should not be deposited in the tissues or provoke functional disturbances; (*d*) they should not agglutinate or hemolyze the erythrocytes; (*e*) they should not alter the osmotic pressure or the ionic or acid-base equilibriums.

Saline solutions. Intravenous injection of saline solution (NaCl, 0.8 to 0.9 per cent; Locke's solution; etc.) can raise the blood pressure in cases of hemorrhage. If the loss of blood is below 2 per cent of the body weight, this treatment may be useful, but if the loss is greater (3 to 5 per cent), if there is a state of shock, or if hypotension has lasted more than 1 hr., whole blood or plasma should be transfused. Saline has only a transitory effect, as the fluid injected passes through the capillaries into the tissues. Subcutaneous or intravenous saline injections are useful in cases of dehydration. It is best to inject them very slowly, drop by drop; an excessively rapid injection can provoke pulmonary edema.

Colloid solutions. The following colloid solutions are used for transfusion: gelatin in concentrations of 5.5 per cent, gum acacia (6 per cent), pectin, methylcellulose, polyvynilpyrrolidone (3.5 per cent), and dextran (6 per cent in saline). Dextran is a polysaccharide of high molecular weight produced from sugar by *Leuconostoc mesenteroides;* it is metabolized in the human body. These substances are called plasma expanders.

Casein hydrolysate has been used for protein feeding in cases in which protein could not be

given by mouth, and also in cases of hypoproteinemia (Elman) (see Chap. 43, Protein Metabolism).

Erythrocyte suspensions. Erythrocytes separated from plasma can be suspended in saline and used for transfusion in cases of anemia or functional disturbances of the red blood cells, as occurs in CO intoxication.

Accidents following Transfusion

Blood transfusion can cause accidents in several ways: (*a*) by an infection carried in the donor's blood; (*b*) by causing gaseous embolism, due to faults in the technique of transfusion; (*c*) by allergic reactions; (*d*) by a too-rapid injection of blood, which can produce dilatation of the right ventricle, acute cardiac insufficiency, and sometimes pulmonary edema; (*e*) by an error in establishing the compatibility of the donor's blood. The last cause is the most frequent; half the deaths following transfusion are due to it.

The effects of transfusion of incompatible blood occur in three stages:

1. Immediately, there is a tingling sensation in the limbs, lumbar pains, precordial oppression, shivering, cyanosis, tachycardia, a fall in blood pressure, and in some cases death. These reactions are due to the obstruction of small blood vessels by the agglutinated erythrocytes.
2. An interval follows during which the clumps of erythrocytes are dissolved, hemoglobin from the hemolyzed erythrocytes diffuses into the plasma, and in some cases jaundice appears.
3. Later there are signs of renal insufficiency, albuminuria, oliguria, and sometimes anuria and death; hemoglobin crystals are found in the renal tubes. In other cases diuresis gradually increases and the patient recovers.

BLOOD GROUPS

Pseudoagglutination and agglutination. When blood in which coagulation is prevented is left standing, the erythrocytes form *rouleaux* and sink to the bottom of the flask; shaking disperses the erythrocytes, which are thus again suspended in the plasma. This formation of *rouleaux* is not real agglutination but pseudoagglutination.

Erythrocytes placed in blood serum taken from a subject of another species are usually agglutinated into clumps that are not dispersed by shaking (heteroagglutination). Erythrocytes placed in blood serum of a subject of the same species can either remain dispersed and gradually settle at the bottom of the receptacle (erythrosedimentation) or become agglutinated (isohemoagglutination, Landsteiner, 1900). In pathologic cases the erythrocytes of an individual can be agglutinated by that individual's own serum (autohemoagglutination).

Blood groups. Any human blood can be classified into one of four groups, according to the presence or absence of two substances in the stroma of the erythrocytes. These substances are known as A and B respectively. They can be found separately or together, or they can both be missing. The absence of both these factors is indicated by the letter O (derived from zero). There are therefore four possible cases: (*a*) the erythrocytes do not have either the A or the B factor (group O); (*b*) the erythrocytes have factor A, but not B (group A); (*c*) the erythrocytes have factor B, but not A (group B); (*d*) the erythrocytes have both A and B (group AB). These substances, which condition isoagglutination, provoke the formation of specific agglutinins when injected into an animal of another species, *e.g.*, a rabbit; the immunologic term for them is "agglutinogens."[1]

Erythrocytes of subjects in the O group are not agglutinated by agglutinins α and β of human sera. Erythrocytes of subjects in group A are agglutinated by the anti-A factor, or α agglutinin. Erythrocytes of subjects in group B are agglutinated by the anti-B factor, or β agglutinin. Erythrocytes of subjects in the AB group are agglutinated by both the α and the β agglutinin. No blood has an agglutinogen in the erythrocytes and the corresponding agglutinin in the plasma.

Landsteiner has established the composition of the blood in the four groups with respect to the agglutinogens in the erythrocytes and the agglutinins in the plasma (Table 11). Thus the α agglutinin (anti-A factor) is found in the plasma of groups O and B, but not in groups A and AB; the β agglutinin (anti-B factor) is found in groups O and A, but not in groups B and AB. Subjects in group O therefore have both the agglutinins in their plasma, and subjects in group AB have neither.

[1] Antigens A and B have been found also in saliva, gastric juice, sperm, certain tissues, etc.

Terminology. The terminology used so far is that proposed by Dungern and Hirzfeld, and adopted by the League of Nations in 1928. The first classifications were those of Jansky and Moss, which used numbers from I to IV; unfortunately, group I in one of these classifications corresponded to group IV in the other. The international classification is simple and logical, as it indicates the agglutinins and agglutinogens in each group.

Table 11. Human Blood Groups

Group	*Erythrocytes*		*Agglutinin in serum*	*International terminology*
	Agglutinogen	*Agglutinated by serum of group*		
O	O	....	α-anti A; β-anti B	$O_{\alpha\beta}$
A	A	O and B	β-anti B	A_{β}
B	B	O and A	α-anti A	B_{α}
AB	A and B	O, A and B	....	AB_o

League of Nations.....	$O_{\alpha\beta}$	A_{β}	B_{α}	AB_o
Jansky (1907).........	I	II	III	IV
Moss (1910)..........	IV	II	III	I

Racial distribution of blood groups. The relative frequency of the blood groups varies in the different human races. It is the same in Western Europe and in the white population of America (Table 12a), but differs in other parts of the world. The proportion of group A diminishes and that of group B increases as one goes from Western Europe toward the East. The maximum frequency of group B is found in the races of the Far East (Hindu, Manchu, and Malay). The Indian tribes of America have a predominance of individuals in group O; in some of them it is nearly 100 per cent. In parts of America where crossbreeding with Indian stock is common, there is a greater proportion of group O than in the rest of the hemisphere.

Table 12a. Percentage Distribution of Blood Groups

Group	*United States*	*Western Europe*	*Buenos Aires**
O	40–45	43	45.6
A	40	40	39.4
B	10–15	13	10.4
AB	5	4	4.6

* Average of 53,584 individuals. The averages for 5,000 individuals in the city of Córdoba were O, 43.2; A, 41.0; B, 12.0; and AB, 3.8.

Inheritance of blood groups. The blood group is inherited following the mendelian laws.[1] Apparently there are three allelomorphs—A, B, and R. Agglutinogens A and B are conditioned by independent and dominant factors, while the absence of agglutinogen (group O) and the factors for agglutinins α and β are recessive.[2] Hereditary transmission of the blood groups is summarized in Table 12b.

Table 12b. Inheritance of Blood Groups

Blood group of parents	*Genetic constitution of parents*		*Blood group to which offspring can belong*	*Blood group to which offspring cannot belong*
O × O	OO	OO	O	A, B, AB
O × A	OO	AA, AO	O, A	B, AB
O × B	OO	BB, BO	O, B	A, AB
O × AB	OO	AB	A, B	O, AB
A × A	AA, AO	AA, AO	O, A	B, AB
A × B	AA, AO	BB, BO	A, A, B, AB	
A × AB	AA, AO	AB	A, B, AB	O
B × B	BB, BO	BB, BO	O, B	A, AB
B × AB	BB, BO	AB	A, B, AB	O
AB × AB	AB	AB	A, B, AB	O

A disputed paternity can sometimes be excluded when the blood groups of the mother, the child, and the putative father are known, but it cannot be definitely affirmed on the evidence of the blood groups. Information obtained by determining to which of the four principal blood groups the parents and offspring belong gives 1 probability in 7 for the exclusion of a mistakenly attributed paternity.

There are two other supplementary factors in the erythrocytes called M and N (see "Subgroups," below). These factors are of importance in blood transfusion only when transfusion is repeated several times, but they are used in cases of disputed paternity; the probability of excluding a wrongfully attributed paternity is thus increased to 1 in 3. More recently several subgroups in group A and an Rh factor have been discovered; using all these tests, wrongfully attributed paternity can be excluded in 45 per cent of the cases.

[1] Dungern, E., and L. von Hirzfeld, *Ztschr. f. Immunitätsforsch. u. exper. Therap.*, **8**, 541–547, 1910.
[2] Bernstein, F., *Klin. Wchnschrft.*, **3**, 1495, 1924, Strandskov, H. H., *Physiol. Rev.*, **24**, 445, 1944.

The blood group is not clearly established in some children until they are over a year old.

Importance of blood grouping in transfusion. When performing a transfusion it is important to bear in mind that the erythrocytes of the donor must not be agglutinated by the

Table 13. Properties of Plasma and Erythrocytes in Each Blood Group

Group	*Plasma or serum*	*Erythrocytes*
O	Agglutinates erythrocytes of groups A, B, and AB	Not agglutinated by serum of any group
A	Agglutinates erythrocytes of groups B and AB	Agglutinated by serum of groups O and B
B	Agglutinates erythrocytes of groups A and AB	Agglutinated by serum of groups O and A
AB	Does not agglutinate erythrocytes of any group	Agglutinated by serum of groups O, A, and B

plasma of the recipient.[1] Blood from a subject belonging to group A_β should not be injected in individuals of groups $O_{\alpha\beta}$ and B_α, as these have α agglutinin in their plasma; it can be injected into individuals of groups A_β and AB_o. Blood of group B should not be injected into individuals of groups $O_{\alpha\beta}$ and A_β, as these have β agglutinin in their plasma; it can be injected into individuals of groups B_α and AB_o. Blood of group $O_{\alpha\beta}$ can be injected into any subject, as whatever the group of the recipient, the erythrocytes will not be agglutinated because they do not carry an agglutinogen; individuals of this group are called "universal donors." Blood of group AB_o can be injected only into individuals of the same group. As these have no agglutinins, they can receive blood from all the groups; they are called "universal recipients."[2]

For all practical purposes, only the agglutinogens in the erythrocytes of the injected blood are of importance. The agglutinins injected are diluted in the recipient's plasma, where there is also an antiagglutinin, so that agglutination of the recipient's erythrocytes does not occur. There are nevertheless some exceptions to this rule: a few individuals have a very high agglutinin concentration, and in certain cases of severe anemia the recipient's erythrocytes can be agglutinated by the donor's agglutinins and cause serious reactions.[1]

[1] Ottenberg's rule.

[2] The terms "universal donor" and "universal recipient" may be dangerous, as sometimes they lead to serious errors. It is best to match the donor's and recipient's bloods before performing a transfusion. Transfusion of blood from a so-called "universal donor" (group O) may be followed by a reaction if its plasma has a high α, β, anti-M, or anti-N agglutinin concentration. Also the donor's blood may have an Rh factor which is absent in the recipient; in this case the recipient will develop antibodies for this factor.

FIG. 16. Agglutination of erythrocytes in the four main blood groups.

Before a transfusion is performed, the donor's and recipient's bloods must be typed. This can be done if active serums from subjects in groups A_β and B_α are kept in stock. A drop of serum A_β and separately another of serum B_α are placed on a slide, and to each drop another of the blood to be tested (preferably citrated to prevent clotting) is added and mixed with the serum. The slides are watched for 20 to 30 min., but

[1] A and B antigens have been prepared (Witebsky) which can be added to blood of group O donors in order to neutralize the α and β agglutinins it contains.

usually agglutination takes place in a shorter time. If clumps are formed only in the drop with serum A_β, the tested blood belongs to group B_α; if they appear in the drop with serum B_α it belongs to group A_β; if agglutination occurs with both serums, the blood is of group AB_o; and if no agglutination occurs with either serum, the blood tested is of group $O_{\alpha\beta}$ (Fig. 16). This method is not infallible, and it is better to match directly the donor's erythrocytes with the recipient's serum.

A cross test between the erythrocytes and serums of the donor and those of the recipient gives reliable information as to the compatibility of the bloods. In a small test tube 0.1 cc. of 0.9 per cent saline solution, 0.1 cc. of a 5 per cent suspension of the donor's erythrocytes, and 0.1 cc. of the recipient's serum are placed. Another test tube receives the recipient's erythrocytes and the donor's serum. Both tubes are left for 20 min. at 37°C. On shaking the tubes the erythrocytes will disperse if agglutination has not taken place but will remain in clumps if it has.

Subgroups. Two subgroups have been found in group A; these are known as A_1 and A_2. A_1 erythrocytes absorb much more α agglutinin than those of the A_2 type. Subgroup A_1 is five to six times more frequent than A_2. There are also subgroups A_1B and A_2B.

Landsteiner and Levine immunized rabbits against human erythrocytes by repeated injections of blood from the same individual. They were thus able to discover two additional factors called M and N. These are found in the erythrocytes of individuals in all the four principal groups, who can have either factor M or N, or M and N together; there are no erythrocytes without one or both of these factors.

Other factors[1] have also been found in the erythrocytes; among these the most important are the Rh and P factors. Taking into account not only the usual groups and subgroups[2] but also the Rh factor, over 2,500 different varieties of blood have been found.

Rh factors. The discovery of the Rh factors made it possible to (*a*) understand and prevent certain accidents caused by the transfusion of blood apparently compatible according to the data given by tests with the A, B, AB, O system;

[1] Schiff and Boyd, 1942.

[2] O, A_1, A_2, A_3 (very rare), B, A_1B, A_2B, A_3B (very rare) M, N_1, N_2, MN_1, P, P_1, P_2, P_1P_2, S, S_1, the different Rh factors, and several others.

(*b*) understand and treat successfully erythroblastosis fetalis (hemolytic disease of the newborn); (*c*) establish more accurately the individuality of the blood, a fact of considerable value in legal medicine and anthropological studies.

Landsteiner and Wiener[1] discovered a factor (which they called the Rh factor) in the erythrocytes of the rhesus monkey, by injecting blood from animals of this species into guinea pigs, thus obtaining a specific immune agglutinin. When human erythrocytes were tested with anti-Rh agglutinin, the Rh factor was found in 85 per cent of the subjects.[2]

Later work showed there were several Rh factors. This was proved by injecting blood into a recipient in order to provoke the formation of the corresponding agglutinin. The anti-Rh standard agglutinin gave 85 per cent of positive cases, and two other agglutinins, anti-Rh_1 and anti-Rh_2, gave 70 and 30 per cent of positive cases respectively. The Rh_1 factor was then found to consist of an Rh and an Rh_0 factor, and Rh_2 of Rh_0 and Rh''. Furthermore, an agglutinin was obtained that gave 80 per cent of positive cases, including all the Rh negatives. This agglutinin was called anti-Hr as these cases appeared to be the reverse of the Rh positive ones.

Fisher and Race[3] observed that a subject could have the Rh' and the Hr factors, or one of these, but could not be lacking in both, a fact that made them suppose these factors were alleles. Rh'' and Rh_0 did not behave as alleles of Rh' or of each other. Fisher and Race therefore stipulated the hypothesis of the existence of three pairs of genes in three loci situated very close to one another in the same chromosome. The Rh' factor was called *C* and its allele Hr, *c*; Rh_0 was called *D* and Rh'', *E*. The existence of alleles *d* and *e* was thus postulated, and these were found later by Mourant[4] in 1945 (factor e) and Haberman *et al.*[5] in 1948 (factor d) which

[1] LANDSTEINER, K., and A. S. WIENER, *Proc. Soc. Exper. Biol. & Med.*, **43**, 223, 1940; **54**, 316, 1943.

[2] This percentage was found in the white races, with the exception of the Basques, of whom only 64 per cent are Rh positive (ETCHEVERRY, M. A., *Ciencia e Investigación*, **3**, 47, 1947). In Negroes, Chinese, and American Indians higher percentages have been found.

[3] RACE, R. R., *Nature*, **153**, 771, 1944.

[4] MOURANT, A. E., *Nature*, **155**, 542, 1945.

[5] HABERMAN, S., J. M. HILL, B. W. EVERIST, and J. W. DAVENPORT, *Blood*, **3**, 682, 1948.

confirmed the Fisher-Race hypothesis. Later a C^w antigen was found in locus C and a D^u antigen in locus D. All the facts so far discovered fit into the three-loci theory of Fisher and Race.

At first the Rh positive cases were considered dominant, and the Rh negatives as recessive, but after the anti-c (anti-Hr) agglutinin was obtained it was shown that both alleles had the same single-dose expression (*i.e.*, were equally dominant), as is the case of the A and B factors in the ABO system considered in a previous paragraph. The other alleles D, d, E, e, also show single-dose expression; therefore whenever any of these factors are genetically present they can also be demonstrated in the erythrocytes by means of the corresponding agglutinin.

Twelve different Rh chromosomes have been identified; they are CDe, cde, cDE, cDe, C^wDe, cdE, Cde, CDE, CD^ue, cD^uE, cD^ue, and C^wde. Theoretically these chromosomes can give rise to 78 different genotypes. Race *et al.*[1] found 13 of these possible genotypes in a series of over one thousand subjects tested (Table 14). There were also two examples of D^u but they are included in the ordinary D groups. Injection of human blood with any type of Rh factor into a subject without this factor in his erythrocytes provokes

Table 14. Frequency of Rh Genotypes

Genotype		*Absolute numbers*	*Percentage*
CDe/CDe	R_1R_1	178	16.59
C^wDe/CDe	$R^w_1R_1$	12	1.12
CDe/cde	R_1r	354	32.99
C^wDe/cde	R^w_1r	9	0.84
CDe/cDE	R_1R_2	138	12.86
C^wDe/cDE	$R^w_1R_2$	6	0.56
cDE/cde	R_2r	137	12.77
cDE/cDE	R_2R_2	29	2.70
cde/cde	rr	170	15.84
cDe/cde	R_0r	19	1.77
Cde/cde	R′r	10	0.93
cdE/cde	R″r	7	0.65
CDe/CDE	R_1R_Z	4	0.37

Source: RACE, R. R., *et al.*, *Blood*, **3**, 689, 1948.

the appearance of the corresponding anti-Rh agglutinin and hemolysin in about 12 days. A second transfusion of the same Rh blood is followed by agglutination and hemolysis of the injected erythrocytes and by severe reactions in the recipient. In these cases blood typing with the α and β agglutinins is not sufficient to establish the compatibility of the donor's and recipient's blood; by repeated transfusion anti-Rh agglutinin is formed in the recipient's plasma, which agglutinates the transfused erythrocytes. Direct matching of the donor's and recipient's blood before each transfusion is the only sure way of knowing if they are compatible.

[1] RACE, R. R., A. E. MOURANT, S. D. LAWLER, and R. SANGER, *Blood*, **3**, 689, 1948.

Hemolytic disease of the newborn. In this disease, also called erythroblastosis fetalis or severe congenital jaundice, the fetus may die *in utero*, with widespread edema and ascites (fetal dropsy), or may be born alive, but with severe hemolytic anemia and jaundice. There are many erythroblasts in the blood, the liver is enlarged, and in some cases there are lesions in the central nervous system. The pathogenesis of the disease was first demonstrated by Levine and his associates, who showed that hemolysis of the fetal blood is due to isoimmunization of the mother by a factor in the erythrocytes of the fetus which provokes the production of antibodies active on the fetal blood. The condition arises when the factor inherited by the fetus from the father is not present in the mother. An Rh factor is involved in the majority (93 per cent) of the cases, but in the remaining 7 per cent the maternal blood is anti-A, or anti-B, etc. Thus if the father is homozygous for one of the Rh factors and the mother Rh-negative for this factor, the offspring will all be Rh-positive for this factor; if the father is heterozygous the offspring may be either positive or negative for this factor. Not all the offspring of Rh-positive fathers from Rh-negative mothers suffer the disease; the disease is seldom, if ever, seen in primiparas, unless the mother has previously received incompatible Rh blood and been immunized by it. Passage of fetal antigen into the mother's circulation apparently does not occur in the course of pregnancy in sufficient amounts to immunize the mother, but it does occur in the course of labor when the placental maternal blood vessels are open; for this reason the mother's blood with successive deliveries becomes increasingly potent against the incompatible fetal factor.

Treatment consists in bleeding the newborn child as soon as possible after birth[1] and replac-

[1] In order to be prepared for emergencies pregnant women's blood should be typed, not only in respect of the A, B, AB, O system, but also in respect of the Rh system, establishing the anti-Rh potency of the serum.

ing its blood by transfusion of blood negative for the corresponding Rh, or other factor, to which the mother has been immunized. Since these erythrocytes are not destroyed by the maternal antibodies, which attack the fetal red blood cells, they tide over the newborn child until these antibodies have been eliminated.

BIBLIOGRAPHY

Hemorrhage

BARCROFT, H., *et al.*, *Lancet*, **246**(1), 489, 1944.

SCHWARTZMAN, G., *et al.*, Hemorrhage, *Ann. New York Acad. Sc.*, **49**, 483–660, 1948.

SHENKIN, H. A., *et al.*, *Am. J. M. Sc.*, **208**, 421, 1944.

WALLACE, J., and E. P. SHARPEY-SCHÄFER, *Lancet*, **241**, 393, 1941.

Blood Groups

HIRZFELD, L., "Les Groupes Sanguins," Masson et Cie, Paris, 1939.

LANDSTEINER, K., *J. A. M. A.*, **103**, 1014, 1934.

LANDSTEINER, K., "The Specificity of Serological Reactions," Charles C Thomas, Springfield, Ill., 1936.

LANDSTEINER, K., *J. Exper. Med.*, **74**, 309, 1941.

LATTES, L., "Individuality of the Blood," Oxford University Press, New York, 1932.

LEVINE, P., A Brief Review of the Newer Blood Factors, *Tr. New York Acad. Sc.*, **13**, 205, 1951.

SCHIFF, F., and W. C. BOYD, "Blood Grouping Technic," Interscience Publishers, Inc., New York, 1942.

WIENER, A. S., "Blood Groups and Blood Transfusion," 3d ed., Charles C Thomas, Springfield, Ill., 1952.

Transfusion

DE GOWIN, E. L., R. C. HARDIN, and J. B. ALSEVER, "Blood Transfusion," Saunders, Philadelphia, 1949.

KEYNES, G., "Blood Transfusion," J. Wright, London, 1949.

KILDUFFE, R. A., and M. DE BAKEY, "The Blood Bank and the Technique and Therapeutics of Transfusion," Mosby, St. Louis, 1942.

MUDD, S., and W. TALHIMER, "Blood Substitutes and Blood Transtusion," Charles C Thomas, Springfield, Ill., 1942.

WIGGERS, C. J., "The Physiology of Shock," Commonwealth Fund, Harvard University Press, Cambridge, Mass., 1950.

Rh Factor

BESSIS, M., "La Maladie hémolytique du nouveau-né," Masson et Cie, Paris, 1947.

CASTLE, W. B., M. M. WINTROBE, and L. H. SNYDER, *Science*, **107**, 27, 1948 (nomenclature).

DAVIDSON, I., P. LEVINE, and A. WIENER, *J.A.M.A.*, **149**, 699, 1952.

DIAMOND, L. K., and F. H. ALLEN, *New England J. Med.*, **241**, 867 and 907, 1949.

LEVINE, P., H. H. STRANDSKOW, A. S. WIENER, and P. VOGEL, *Bull. New York Acad. Med.*, **25**, 244, 1949.

MOLLISON, P. L., A. E. MOURANT, and R. R. RACE, The Rh Blood Groups and Their Clinical Effect, Med Res. Council, Mem. No. 27, London, 1952.

POTTER, E. L., "Rh. Its Relations to Congenital Hemolytic Disease and to Intragroup Transfusion Reactions," Year Book Publishers, Inc., Chicago, 1947.

RACE, R. R., and R. SANGER, "Blood Groups in Man," Blackwell, Oxford, 1950; Charles C Thomas, Springfield, Ill., 1951.

WIENER, A. S., *Brit. M. J.*, No. 4657, 1950; *Bull. W. H. O.*, **3**, 265, 1950; *Bull. Pan-American Sanit. Bureau*, **30**, 307, 1951 (heredity, paternity, nomenclature).

CHAPTER 8

Coagulation of the Blood

Description. Blood loses its fluidity when it is shed outside the blood vessels; at first its viscosity increases, and then it solidifies into a jelly, so that the vessel which contains it can be inverted without spilling it. A short time after the blood has coagulated, the clot contracts and a yellow fluid (the blood serum) is squeezed out.

It is the plasma that clots. Fibrinogen is found in the plasma in the state of hydrosol; on coagulation it is transformed into a filamented hydrogel called fibrin. Beside blood, other fluids in the organism which contain fibrinogen clot spontaneously when they are shed, *e.g.*, lymph and pathologic exudations. Some other fluids containing fibrinogen, such as normal pericardial and pleural fluid, the liquid in a hydrocele, and pathologic transudations, do not clot spontaneously.

Microscopic observation shows that the clot is formed by a network of fine filaments, which entangles the red and white blood cells in its meshes; the platelets adhere to the filaments once these have been formed. Ultramicroscopic observation shows the formation first of pseudocrystalline needles, which then form threads and the fibrin network. Sometimes blood clotting is incomplete; only clumps of fibrin are formed, but the whole mass of blood does not become solid. If recently shed blood is shaken with glass beads or beaten with a bundle of fine rods while it is clotting, fibrin will be deposited on the beads or rods in the shape of white elastic filaments, to which fibrin owes its name; the remaining fluid, called defibrinated blood, is a suspension of blood cells in serum.

Physiological significance of coagulation. Coagulation of the blood helps to stop hemorrhage, because the clot closes the opening in the blood vessel. When there is a deficiency in blood coagulation, *e.g.*, in hemophilia, a slight accidental wound or a small surgical operation, such as the extraction of a tooth, can cause prolonged bleeding, which lasts for hours or even days and can imperil the patient's life. Normal clotting of shed blood is therefore a defensive mechanism. On the other hand clotting can occur within a blood vessel (thrombosis) and, by obstructing the vessel, cause necrosis in the tissues in which the circulation has been stopped. In other cases an intravascular clot is set free from the place where it has been formed and is carried by the blood stream to a distant part of the body where it obstructs a blood vessel (embolism); this may have serious and sometimes fatal results.

Contraction of the clot and fibrinolysis. Contraction of the clot is similar to syneresis in colloids. Sometimes it takes place rapidly and is marked; at other times it does not occur. Contraction of the clot is conditioned by the concentration of thrombin and thrombocytes. The importance of the latter is shown by the following facts: (*a*) contraction is increased and accelerated by addition of thrombocytes; (*b*) it does not occur in cases of thrombopenia (less than 60,000 thrombocytes per cubic millimeter) or when antithrombocyte serum is added (Le Sourd and Pagniez).

Fibrin is usually slowly dissolved in serum (fibrinolysis). Sometimes fibrinolysis takes place rapidly, *e.g.*, in surgical and traumatic shock, in chloroform intoxication, and in shock produced by the injection of peptone. It is provoked by an enzyme, fibrinolysin (plasmin), which acts as a protease, *i.e.*, a proteolytic enzyme, digesting not only fibrin, but also fibrinogen and prothrombin. In normal plasma, there is profibrinolysin (plasminogen) which is activated by a coenzyme,

fibrinokinase, and is converted into fibrinolysin. Other factors also activate profibrinolysin, *e.g.*, tissue extracts, the streptokinase in streptococci, and substances released by shock (traumatic, peptonic, etc.). An antifibrinolysin has been found, which inhibits fibrinolysin.

Role of the tissues. Blood taken directly from a blood vessel clots less rapidly than blood that has been in contact with the tissues. Damaged cells release substances which accelerate clotting. This phenomenon can be easily demonstrated in birds. Blood which is shed through a wound clots within a few minutes, but if it is drawn directly from an artery and immediately centrifuged, the plasma remains fluid indefinitely. This can also be done with mammalian blood, but it must be drawn into tubes coated with silicone in order to avoid damage to the thrombocytes (which are very fragile) so that by ultracentrifugation they can be eliminated without having been damaged.

THE MECHANISM OF BLOOD COAGULATION

Many theories have been proposed to explain the process of blood coagulation, and a complicated, profuse terminology has been used. This complexity is due to the facts that the knowledge of the mechanism of blood clotting is still incomplete and that most of the substances that play a part have not yet been obtained as pure chemical entities.

There is almost universal agreement as to the transformation of fibrinogen into fibrin by the action of thrombin. The latter does not exist in the circulating blood but is rapidly formed in shed blood by the action on prothrombin of thromboplastin, calcium ion, and an activating or accelerating factor.

The following diagram summarizes this process:

Circulating blood: Thromboplastinogen; Thrombocytes; Thromboplastin cofactor AC globulin (inactive); Ca^{++}; Prothrombinogen → Prothrombin; Fibrinogen; Profibrinolysin

Shed blood — First stage: Thrombotonin (vasoconstriction) ← Thrombocytes; Thrombocytes → Thromboplastinogenase; Thromboplastinogen ← Thromboplastinogenase; Thromboplastinogen → Thromboplastin; Ca^{++}; Thromboplastin cofactor AC globulin (active, labile); Prothrombin; Thromboplastin → (prothrombin complex) → Thrombin

Second stage: Antithrombin (plasma) → Thrombin → Fibrinogen; Thrombin → Metathrombin; Fibrinogen → Fibrin; Profibrinolysin ← Fibrinolysinokinase; Antifibrinolysinokinase → Fibrinolysinokinase; Profibrinolysin → Fibrinolysin; Thrombocytes (contraction of clot) → Fibrin ← Fibrinolysin; Antifibrinolysin → Fibrinolysin; Fibrin → Fibrin-derivates

Heavy line: Quick's prothrombinogenic cycle

FIBRINOGEN

Fibrinogen is the substance that clots and forms fibrin; therefore, only those fluids which contain it can coagulate. It is a globulin with a large molecule; its molecular weight is 500,000, and it has a high viscosity. It is very unstable in the presence of salts. It coagulates when heated to 56 to 60°C. for 10 min. It can be prepared by precipitating citrated plasma with an equal volume of saturated NaCl solution. The precipitate is redissolved by dilution (preferably with citrate) and reprecipitated at least three times, to purify it. Recently it has been obtained by precipitation with ethanol at 0°C. Clotting of fibrinogen is caused by the addition of thrombin, but not by the addition of any one of the substances that give rise to

thrombin, *i.e.*, prothrombin, thromboplastin, or Ca^{++}. Plasma contains only 0.25 gm. per cent of fibrinogen, yet this amount is sufficient to solidify the whole blood, because it forms a network that entangles the cells and absorbs water.

Fibrinogen is formed in the liver (perhaps exclusively). After hepatectomy the concentration of fibrinogen in plasma remains normal or falls slightly; but if blood is repeatedly drawn from a hepatectomized animal, defibrinated, and reinjected until there is no fibrinogen in the plasma, fibrinogen is not regenerated. The small amount of fibrinogen which is replaced apparently comes from stores in the bone marrow and the intestine. If the same procedure is carried out in a normal animal, fibrinogen is regenerated in a few hours.

Thrombin[1]

Thrombin coagulates solutions of pure fibrinogen and oxalated or citrated plasma if there is an adequate electrolyte balance. Thrombin is the cause, not the effect, of coagulation; it is formed from prothrombin by the action of thromboplastin, Ca ion, and a cofactor for thromboplastin even when there is no fibrinogen and therefore fibrin cannot be produced. Thrombin is one of the euglobulins and has carbohydrate groups.[2] When the blood clots, the greater part of the thrombin is found in the serum. Thrombin is adsorbed by fibrin and other surfaces electrically charged (*e.g.*, glass), but not by those that have no charge, *e.g.*, vascular endothelium, surfaces coated with silicone or paraffin, etc. Fresh serum contains a large amount of thrombin, which can be precipitated by alcohol and stored in powdered form; when this powder is dissolved in water, thrombin is set free. There is no thrombin in circulating blood. This fact can be proved by drawing blood into alcohol; the precipitate does not contain thrombin. Because of this lack of thrombin, circulating blood does not clot.

There has been much discussion as to whether thrombin acts as an enzyme or enters into a chemical combination. The following facts have been advanced in favor of an enzymatic activity:

1. At 0°C. thrombin is inactive. Its activity increases with temperature and starts to decline at 50°C. At this temperature it is partially destroyed, and after 5 min. at 60°C. it is completely destroyed.
2. The clotting time is shortened as the concentration of thrombin increases.
3. Thrombin is not used up in the course of coagulation; on the contrary, it increases.

Thrombin is spontaneously inactivated in serum by antithrombin; at 35 to 50°C. inactivation is rapid. An inactive metathrombin is formed, which can be reactivated by adding alkali.

Some snake venoms have thrombins which differ somewhat from those in serum, but which clot fibrinogen solutions, oxalated plasma, and blood.

[1] Fibrinferment (Schmidt), thrombase (Mellanby), plasmase (Duclaux).
[2] Seegers, W. H., *J. Biol. Chem.*, **136**, 103, 1940; **146**, 511, 1942; Robbins, K. C., *Arch. Biochem.*, **6**, 69, 1945.

Calcium Ion

Arthus and Pagés[1] showed that certain anions, such as oxalate and citrate, whose calcium salts have a low solubility, inhibit blood clotting. Restitution of calcium provokes clotting. Calcium acts when in the ionized state (Sabbatani). Furthermore, the addition of salts in high concentration, *e.g.*, NaCl, Na_2SO_4, or even $CaCl_2$, prevents clotting (Schmidt). Coagulation does not take place when the ionic strength of the blood (or plasma) is too high (0.5) or too low (0.01). The concentration of calcium necessary to provoke clotting of decalcified plasma is at a minimum when the ionic strength is 0.03. It must be increased proportionally as the ionic strength is raised. The optimum proportion of Ca^{++} to other ions is 1:16. Other alkaline earths (Mg, Ba) can substitute for calcium, but they act only at higher concentrations; the relative concentrations of Ca^{++}, Mg^{++}, Ba^{++} needed to provoke clotting of decalcified plasma are 1:8:60. Anions inhibit coagulation by a double mechanism: (*a*) by lowering calcium ion concentration (decalcifying effect); (*b*) by raising the total ionic strength. Citrate is more effective than fluoride because of its higher valency (ionic strength); oxalate is approximately as effective as citrate because its lower valency is compensated for by its superior decalcifying effect.[2]

[1] Arthus, M., and C. Pagés, *Arch. de physiol. norm et path.*, **2**, 739, 1890.
[2] Lovelock, J. E., and B. M. Porterfield, *Biochem. J.*, **50**, 415, 1952.

How calcium acts and why it is more effective than other bivalent cations is not well understood. According to Morawitz Ca^{++} has a catalytic effect. It has been suggested that calcium combines with prothrombin and gives an unionized complex,[1] but this theory is not supported by part of the evidence available. It has also been suggested that calcium neutralizes a coagulation inhibitor.[2] The greater effectiveness of calcium may be due to its special solubility in the blood; thus a saturated solution of calcium in plasma ultrafiltrate has twice the calcium ion found in a saturated solution of Ca in distilled water. Calcium ion can neutralize the charge of colloids, *e.g.*, of phosphate colloids. The relative effectiveness of Ca^{++}, Mg^{++}, and Ba^{++} in neutralizing the charge of sphingomyelin is similar to that found in blood clotting; moreover the concentration of the cations must be raised as the total ionic strength increases, in the same way as in coagulation. Calcium plays its part mainly in the first phase of blood clotting, but it also acts in the second phase, *i.e.*, in the conversion of fibrinogen into fibrin by the action of thrombin.[3]

Coagulation is, therefore, dependent on the electrolyte balance of the medium, which conditions the interaction of plasmic colloids, and calcium plays a special part in the maintenance of this balance.

Prothrombin[4]

Prothrombin is a glycoprotein which has been obtained in purified form.[5] There is 15 to 20 mg. of prothrombin per 100 ml. of plasma. It is destroyed by acids, by alkalies, and by heating at 62°C. It is easily adsorbed by tricalcium phosphate, magnesium hydroxide, etc., and can be separated again by elusion in sodium citrate solution. Plasma can be completely deprived of prothrombin by adsorption.

Pure prothrombin does not coagulate fibrinogen. A 25 per cent solution in sodium citrate becomes activated spontaneously (autocatalysis) in 16 hr. (Seegers). In the presence of thromboplastin, cofactor (Quick's labile factor), and calcium ion it is rapidly converted into thrombin. The velocity of this conversion is dependent on these three factors. Certain substances such as gum acacia, gelatine, fine powders, etc., are known as thromboplastic or fibrinoplastic agents because, though not essential to the process, they accelerate it. Activated prothrombin coagulates plasma very rapidly; 1 mg. will clot 2 liters of plasma in 15 sec.

According to several investigators, prothrombin is found in normal circulating plasma in the form of a precursor, called proserozyme (Bordet and Delange) or prothrombinogen (Quick), which is converted into the prothrombin of serum.

Prothrombin is formed in the liver if there is a requisite amount of vitamin K. Prothrombin insufficiency may arise owing to mechanical, thermal, or chemical damage to the liver, or to vitamin K deficiency; also a rare form of congenital prothrombin insufficiency has been described (Quick).

Intoxication by dicumarol inhibits prothrombin formation in the liver.

Vitamin K

Vitamin K (Dam's coagulation vitamin) is necessary for the normal production of prothrombin. When it is not found in the diet or is not synthesized by bacteria in the intestine, prothrombin in the plasma diminishes and spontaneous hemorrhage occurs easily. It is dangerous to submit a patient in this condition to any surgical operation.

Vitamin K was discovered by Dam (1929) in chicks fed a special diet, which provoked a severe condition with ulcers in the gizzard and incoagulability of the blood. These disorders were cured by giving a substance called vitamin K, found in alfalfa, spinach, putrified fish flour, several bacteria, etc. Two chemical substances were isolated which had vitamin K activity (Doisy): vitamin K_1, or 2-methyl-3-phytyl-1,4-naphthoquinone, extracted from alfalfa, and later prepared by synthesis; and vitamin K_2, 2,3-difarsenyl-1,4-naphthoquinone, extracted from putrified fish flour. Several substances prepared by synthesis have the same activity as the natural vitamins K, although their chemical

[1] Quick, A. J., and M. Stephanini, *J. Gen. Physiol.*, **32**, 191, 1948.

[2] Overman, R. J., "Blood Clotting and Allied Problems," Josiah Macy, Jr., Foundation, New York, 1949.

[3] Fibrinogen solution purified by electrodialysis is not clotted by thrombin also purified by electrodialysis, but it does clot when a small amount of calcium is added (Rabinovich, R., *Combt. rend. Soc. de biol.*, **95**, 755, 1926).

[4] Prothrombin (Schmidt, Mellanby), serozyme (Bordet), plasmozyme (Spiro), thrombogen (Nolf).

[5] Seegers, W. H., *et al.*, *Arch. Biochem.*, **6**, 85, 1945; *Am. J. Physiol.*, **148**, 563, 1947; **150**, 58, 1947; Robbins, *op. cit.*, p. 75.

structure is simpler. The most active of these is 2-methyl-1,4-naphthoquinone (menadione); it has two to four times the potency of the natural vitamin K.

The maintenance of a normal concentration of prothrombin in blood plasma requires the absorption of a certain amount of vitamin K. This vitamin is not only found in the diet but also synthesized by the bacteria in the intestine. Bile is necessary for the absorption of vitamin K. There is a tendency to spontaneous hemorrhage when the prothrombin in plasma falls below 20 per cent of the normal concentration, and in some cases with an even less marked decrease. Prothrombin diminishes in the following conditions: (*a*) when there is no vitamin K in the diet; (*b*) when vitamin K is badly absorbed because there is no bile in the intestine (obstruction of the common bile duct, biliary fistula); (*c*) in intestinal diseases which disturb absorption (sprue, ulcerative colitis); (*d*) in severe hepatic lesions (hepatosis, intoxications which damage the liver); (*e*) in certain intoxications, such as those following administration of dicumarol, and sometimes of salicylate or sulfaguanidine. Natural or synthetic vitamin K given by mouth, especially when bile is also given so as to facilitate its absorption, improves blood clotting in all these conditions except (*d*) and (*e*).

In the newborn child there is a low concentration of prothrombin in the plasma, which can be the cause of severe hemorrhage. It is therefore advisable to give vitamin K to the mother before delivery, or to the child if there are signs of retarded blood coagulation.

In some animals intestinal synthesis by bacteria is an important source of vitamin K.

Thromboplastin[1]

Thromboplastin rapidly converts prothrombin into thrombin in the presence of calcium ion. This process is considerably accelerated by an activating cofactor (Quick's labile factor). Thromboplastin alone does not clot solutions of pure fibrinogen, or oxalated plasma.

Thromboplastin is found in plasma, especially after disintegration of the thrombocytes. It is also found in human milk and in tissue extracts, especially those of brain, lung, thymus, and placenta. Aqueous extracts are thermolabile, and alcoholic extract with the tissues clots more rapidly than blood drawn directly from a blood vessel; this has been attributed to the effect of tissue thromboplastin. There seem to be two groups of active substances: (*a*) phospholipids related to cephalin; (*b*) lipoproteins not yet well identified.

Quick and others maintain that thromboplastin is originated in a precursor called thromboplastinogen, which is activated by an enzyme in the thrombocytes (thromboplastinogenase); thromboplastin deficiency in cases of thrombocytopenia would therefore be due to a deficiency of this enzyme. According to Ferguson a protease (tryptase) activates thromboplastin and calcium. Inhibitors of thromboplastin have been described (Morawitz, Tocantins). It has been postulated that the passage of placental thromboplastin into the maternal circulation is in part responsible for the condition known as toxemia of pregnancy.

Thromboplastin Accelerator or Cofactor[1]

The conversion of prothrombin into thrombin by thromboplastin and calcium ion is considerably accelerated by a factor in the absence of which the activity of thromboplastin is markedly retarded and diminished. Deficiency of this factor is the cause of retarded clotting and prolonged hemorrhage in certain cases (Owren, Seegers). It is found in very small quantities in normal plasma and is activated by the thrombocytes.

The accelerator factor rapidly diminishes in stored plasma after 1 week, although prothrombin remains unaltered. Hepatic lesions cause both substances to diminish. Dicumarol provokes a lasting decrease in prothrombin, but only a transitory decrease in accelerator factor. Aminophylline and methylxanthines increase thromboplastin cofactor.

According to Seegers prothrombin, thromboplastin, calcium ion, and an accelerator in the thrombocytes form a small amount of thrombin. Thrombin converts inactive AC globulin in plasma into the active AC globulin found in serum. Thrombin would then be formed very rapidly by the interaction of prothrombin,

[1] Thromboplastic or zymoplastic factor (Schmidt), thrombokinase (Morawitz), cytozyme (Bordet, Spiro), thrombozyme (Nolf).

[1] Factor V or convertin (Owren); prothrombin A or labile factor (Quick), accelerator factor (Fantl and Nance), cofactor of thromboplastin (Honorato); accelerator globulin or AC globulin (Ware, Guest, and Seegers).

thromboplastin, calcium ion, active AC globulin of serum, and the accelerator in the thrombocytes.

Owren maintains that thromboplastin with a factor called proconvertin forms convertin, which in the presence of calcium ion converts prothrombin into thrombin. Thrombin thus formed activates proaccelerin (factor V) into accelerin (factor VI) which has a strong accelerator effect on the conversion of prothrombin into thrombin.

The Role of the Thrombocytes

Thrombocytes play an important part in blood clotting. Body fluids that contain fibrinogen but do not clot spontaneously usually coagulate when thrombocytes are added. Any factor which facilitates the adhesion and disintegration of thrombocytes, such as contact, shaking, fine powders, alcohol, or chloroform, causes or accelerates clotting. On the other hand, clotting can be retarded or prevented by drawing blood directly from a blood vessel through a needle coated with silicone into a receptacle also coated with silicone so that the thrombocytes do not adhere to the surfaces with which they come into contact and do not disintegrate. In cases of thrombopenia the blood clots slowly.

Thrombocytes were thought to act on blood clotting by the release of thromboplastin, but it seems that they contain very little thromboplastin. Thrombocytes take part in the formation of thrombin by means of the accelerator factor for thromboplastin (Seegers) and of thromboplastinogenase (Quick) which converts inactive thromboplastinogen into active thromboplastin.

Thrombocytes are an important factor in the retraction of the clot. Also when they accumulate in an opening on the blood-vessel wall they release a vasoconstrictor substance (thrombotonin) which reduces the lumen of the damaged blood vessel, thus contributing to stop the loss of blood.

Theories on the Mechanism of Blood Clotting

As we have seen, a number of factors take part in blood clotting. As yet it has not been possible to experiment with them in their pure chemical form. This limitation has led to difficulties in the interpretation of even well-known facts and has brought forth a great number of theories. These can be divided into physical and chemical theories, although both physical and chemical phenomena play a part in the complex process of coagulation.

Most of the chemical theories are gradually being fused into a common one. There is general acceptance that in shed blood thrombin is formed and acts on fibrinogen, which is transformed into fibrin. This is the fundamental process of coagulation. There are two stages: in the first, thrombin is formed; in the second, thrombin converts fibrinogen into fibrin. Differences in the interpretation of the facts refer to the first phase.

Present theories on the mechanism of blood derive from others submitted some years ago, especially those of Morawitz, and Fuld and Spiro. Bordet, Quick, and others distinguish three instead of two stages in the process of coagulation. In the first stage, thromboplastin and prothrombin are formed from inactive precursors; in the second, the interaction of prothrombin, thromboplastin and its cofactor, and calcium ion forms thrombin; in the third, thrombin converts fibrinogen into fibrin.

There are several physical theories. Some observers maintain that coagulation is due to the union of several colloids in the plasma (Wooldridge, Nolf, Mills) and that thrombin is an effect, not the cause, of coagulation. According to Nolf, thrombozyme (prothrombin) + thrombogen (thromboplastin) + Ca + fibrinogen → fibrin + thrombin. Other physical theories have been advanced (Hekma, Stuber).

ANTICOAGULATING FACTORS

Antithrombin, heparin, and antithromboplastin are factors which inhibit clotting in the organism. In therapeutics heparin and dicumarol or its derivatives are used.

Antithrombin. Old plasma or serum and a few defatted tissues prevent or retard the action of thrombin or fibrinogen. This property has been located in the albumin fraction of serum and plasma. It is attributed to a factor called antithrombin which so far has not been obtained in pure form.

Heparin. Heparin was first extracted from the liver (McLean, 1916) and called heparin by Howell, although it can also be extracted from the lung, the gut, and other tissues. It exerts an immediate anticoagulating effect *in vitro* and *in vivo*, but it must be infused continually or injected repeatedly in order to maintain the blood

incoagulable. There is little or no heparin in normal blood; therefore it is not antithrombin. It does not inhibit the effect of thrombin on fibrinogen unless a factor found in serum albumin (heparin complement) is added. Seegers maintains that thrombin is not destroyed, but that its adsorption by fibrin is increased so that it cannot act on fibrinogen. Heparin has been attributed antiprothrombin activity; also it seems to inhibit or retard the adhesion of thrombocytes on a glass surface (Best *et al.*). It is found in the mast cells of the liver and lung and in the basophil leukocytes of the blood (Wilander). It appears in appreciable quantities in the blood in the course of anaphylactic or peptonic shock and following irradiation in dogs with a hemorrhagic syndrome.

Purified heparin is used in therapeutics to prevent clotting in blood transfusion, to prevent the formation of thrombi after operations or in other situations in which there is danger of thrombosis, in vascular surgery, to prevent the formation of pleural adhesions, etc.

The barium salt of heparin has been crystallized (Charles and Scott). Heparin is mucoitin-polysulfuric acid. Mucoitin is a compound of glycuronic acid and an acetyl derivative of glycosamine. Heparin can be demonstrated in blood and in tissues because it changes the color of toluylene blue to purple. The activity of heparin is measured by comparing its anticoagulating effect with that of an international standard heparin powder, which has 130 units per mg. Heparin can also be estimated by the amount of prothrombin that must be added to neutralize its anticoagulating effect.

Heparin contains 13.8 per cent of S in the form of SO_4. Several anticoagulants have SO_4 in their molecule, and anticoagulating activity is developed in certain substances such as cellulose, starch, and glycogen by incorporating SO_4 into their molecule. One of these, paritol, is used in human therapeutics.

Antithromboplastin. A factor which inhibits thromboplastin has been described by Morawitz, Tocantins, and others. Its chemical constitution and physiological significance are not yet well known.

Dicumarol. A toxic substance, 3,3′-methylene-bis (4 hydroxycoumarin), called dicumarol, has been obtained from fermented sweet clover (*Melitotus* sp.) which provokes hemorrhagic disease in cattle. It has also been prepared by synthesis. It acts on the liver, inhibiting the formation of prothrombin, so that the blood becomes incoagulable. This is not an immediate effect like that of heparin, but takes 1 to 2 days to develop. Prothrombin remains low for a long time, and there is also a transitory decrease in thromboplastin accelerator (AC globulin). Dicumarol activity is measured by determining the prothrombin time, which is prolonged. It has been used in order to prevent postoperative intravascular clotting and to reduce the growth of thrombi, especially in coronary thrombosis.

Hypoprothrombinemia with retarded blood clotting has been reported in cases of intoxication by salicylate or sulfaguanidine (see "Vitamin K" in Chap. 49).

THE FLUIDITY OF THE BLOOD

Circulating blood does not clot because it does not contain thrombin. Moreover prothrombin, thromboplastin, and its accelerator seem to be present as inactive precursors; the thrombocytes are also intact. The stability of the thrombocytes and of the fluid state is dependent on the integrity of the vascular endothelium to which the thrombocytes do not adhere. This is well demonstrated in the following experiment: The jugular vein, filled with blood, is separated from the body between two ligatures. The blood can remain fluid for a long time, but if the endothelium of the vein is damaged or a thread is passed through the vessel, a clot is formed on the damaged surface and all the blood in the vein coagulates. Blood will circulate indefinitely through an anastomosis between two vessels if an intact endothelial surface is carefully maintained. Damage to the endothelium causes local formation of a clot on the wall or within the blood vessel, as is seen in aneurysms and in thrombosis.

There is antithrombin activity in circulating plasma sufficient to neutralize very small quantities of thrombin; injection of a large dose of thrombin provokes massive coagulation of all the blood. An antithrombokinase (antithromboplastin) activity has been described (Morawitz, Tocantins), but its significance is doubtful. The presence of a stabilizing colloid has been postulated (Pickering), but there is no final proof of its existence. Some snake venoms provoke slow intravascular clotting, without the formation of large clots. Fibrinogen is thus used up, and the blood becomes incoagulable; it

has become defibrinated blood (Mellanby, Houssay and Sordelli).

Blood shed into the pleural cavity in moderate quantities remains fluid. It has little fibrinogen and has antithrombin activity; possibly clotting takes place very slowly, and the small amount of fibrin formed is dissolved by fibrinolysis. Menstrual blood remains fluid, except when there is an abundant flow. It has no fibrinogen or thrombin. Probably it coagulates but the clot is dissolved, as a fibrinolytic enzyme has been found (Smith). There is a toxic euglobulin (necrosin) in the menstrual flow.

AGENTS THAT MODIFY COAGULATION

Anticoagulants in Vitro

The following agents prevent or retard coagulation of shed blood:

Cold retards but does not prevent clotting. Local application of cold to a wound can nevertheless stop bleeding by provoking vasoconstriction. Irrigation with warm saline can also stop hemorrhage in some cases by provoking contraction of smooth muscles, *e.g.*, in the uterus.

Prevention of contact. When the blood comes into contact with a foreign surface, thromboplastin and prothrombin are formed from their precursors and thromboplastin accelerator is activated. This contact can be avoided by coating the cannulas and vessels into which the blood is drawn with paraffin or another substance such as silicone that prevents the blood from "wetting" the surfaces with which it comes into contact. Lusteroid tubes also serve this purpose.

Decalcifying agents. Any substance that combines with Ca, so that it is no longer a free ion, prevents blood coagulation. Sodium or potassium oxalate (0.1 per cent), sodium citrate (0.2 to 0.3 per cent), sodium fluoride (0.3 to 0.5 per cent)[1] are commonly used in the laboratory as anticoagulants; sodium citrate, because of its innocuity, is used in medical practice to maintain the fluidity of the blood for transfusion. More recently, ethylene diamine tetra acetic acid is used as an agent capable of chelating calcium ions and maintaining blood fluidity and integrity of platelets.

Concentrated salts. Blood collected into a concentrated salt solution does not clot. The following salts have been used for this purpose: $MgSO_4$ (1:4 of a 25 per cent solution); Na_2SO_4 (an equal volume of a half-saturated solution); NaCl (1:2 of a 10 per cent solution). These salts prevent the formation of thrombin. If the blood is centrifuged or left standing for a sufficiently long time, the plasma is freed from platelets and therefore of thromboplastin, so that it does not coagulate when it is diluted.

Antithrombins. The salivary glands of the leech and the digestive tract of other hematophagous animals secrete anticoagulating substances. Hirudin (Haycraft 1884), an extract of leech's head, and novhirudin, a more purified preparation, have antithrombin activity, and at one time they were frequently used in the laboratory. Heparin (1 mg. for 3 to 5 cc. of blood) has now replaced antithrombins of animal origin in both laboratory and medical practice.

Snake venoms. Some of the snake venoms act as anticoagulants by one of two mechanisms: (*a*) a phosphatidase destroys thromboplastin, as in the case of the venom of the hooded cobra, *Naja tripudians;* (*b*) fibrinogen is attacked so that it no longer clots, as with the venom of *Lachesis flavoviridis.*

Other anticoagulants. Several dyes prevent clotting; many of them have sulfon groups in their molecule, *e.g.*, sodium polyanetholsulfonate (Liquoid Roche), Chicago blue 6B, chlorazol fast pink, etc.

Agents That Accelerate Coagulation in Vitro

The following agents accelerate the coagulation of shed blood:

Heat. At 37°C. blood clots faster than at lower temperatures.

Mechanical contact. Gentle shaking accelerates clotting; vigorous shaking retards it, because it breaks the network of fibrin while it is forming. Clotting is also accelerated by increasing surface contact, *e.g.*, by introducing threads or cotton into the blood, or adding certain colloids.

Calcium. A slight increase in the blood calcium ion concentration accelerates clotting; a high concentration of calcium salts retards clotting.

Thromboplastin. The addition of thromboplastin shortens the coagulation time considerably. Thromboplastin is spread on bleeding surfaces to accelerate clotting and thus stop the

[1] The figures given are those for man; for the dog they should be doubled.

hemorrhage. Fresh aqueous extract of brain, alcoholic extracts of platelets or tissues, and the diluted venom of certain snakes (*Vipera russelli*) are used for this purpose.

Thrombins. Thrombin activity is found in (*a*) fresh serum; (*b*) thrombin extracted from clotted blood; (*c*) some snake venoms (*Bothrops*). The antithrombin in normal blood has very little activity against the thrombin of snake venoms.

Changes in Blood Coagulability in Vivo[1]

A decrease in fibrinogen retards the clotting of blood, and when all the fibrinogen is lost, the blood remains permanently fluid. This is observed (*a*) in certain degenerative processes of the liver; (*b*) in rare cases of congenital hypofibrinogenia, in which fibrinogen is formed in subnormal amounts, or of afibrinogenia, in which it is not formed at all; (*c*) in persons bitten by snakes with coagulating venoms, such as that of *Bothrops* (defibrination of the blood).

The injection of several substances into the circulation of an animal can render its blood incoagulable.

These substances can be divided into two classes:

1. Those which modify the coagulability of the blood as they do *in vitro*, without provoking an organic reaction. Among these substances, heparin should be mentioned in the first place because it is commonly used in therapeutics to prevent intravascular clotting, as it is well tolerated by the organism. Hirudin, polyanethol sulfonate (Liquoid Roche), and anticoagulant dyes are less used in laboratory experiments. Cobra venom and decalcifying salts are used only for experimental purposes because of their toxic effects.
2. Those which provoke an organic reaction and thus modify the coagulability of the blood. Among the substances of this class, the following can be mentioned: (*a*) peptone and substances that provoke anaphylactic shock; (*b*) tissue extracts; (*c*) snake venoms; (*d*) dicumarol.

Peptone shock, produced by rapid intravenous injection of peptone, and anaphylactic shock produce a typical triad: (*a*) hypotension; (*b*) leukopenia; (*c*) incoagulability of the blood. Hypotension is due mainly to contraction of the suprahepatic veins, which causes stasis in the portal system and thus diminishes the venous return; there is also some peripheral vasodilatation. Leukopenia is caused by the retention of the leukocytes in the abdominal blood vessels. Incoagulability of the blood is produced by the liberation of heparin from the liver; heparin prevents the blood from clotting even when thrombin is added. In hepatectomized animals, peptone does not provoke incoagulability of the blood.

The injection of tissue extracts or of thrombin produces different effects according to the dose:

1. Large doses provoke massive clotting of the blood.
2. Small doses produce shock, with hypotension and leukopenia. In a first stage blood clotting is accelerated (positive phase); in a later stage it is retarded or prevented (negative phase). This negative phase is due to the antithrombin effect produced by the discharge of heparin from the liver.

Injection of thrombin extracted from clotted blood, or of thrombin-acting snake venoms[1] is followed by a positive and a negative phase. The negative phase is accompanied by a decrease in the concentration of fibrinogen or even the total absence of fibrinogen from the plasma (defibrinated blood).

Dicumarol retards coagulation because it diminishes prothrombin. The effects are evident only after one or two days, and they last from 2 days to 2 weeks, according to the dose.

TECHNIQUES FOR DETERMINING THE COAGULABILITY OF THE BLOOD

Coagulation time. Many different methods have been proposed; in all of them the time in which the blood clots in certain standard conditions is determined, and compared with the normal. The end point is detected in different ways, and considerable experience is required to obtain comparable results. The usual procedure is as follows: Blood is drawn from a vein into Pyrex glass tubes 8 mm. in diameter, previously rinsed with isotonic salt solution; 1 cc. of blood is put into each of three tubes, and these are submerged in a bath at 37°C.; every 30 sec. one or two of the tubes are examined, and by gently inclining them the formation of the clot is observed; the end point is taken when the

[1] Quick, A. J., *Physiol. Rev.*, 24, 297, 1944.

[1] The venoms of *Bothrops alternata* (*yarará* snake) and *B. neuwiedii* also provoke fragility of the blood vessels and hemorrhages in the nose, gums, and intestinal tract.

clot adheres to the walls of the tube. The third tube is left quiet and examined only when the blood in the others has clotted. The average time taken by the three tubes is the coagulation time; normally it varies from 5 to 15 min.

Bleeding time. A small wound is made by pricking the lobe of the ear with a lancet or a needle; the blood that flows is dried by gently touching the wound with filter paper every 30 sec.; when the paper is no longer stained, bleeding has ceased (Duke). Normal subjects have a bleeding time of 1 to 4 min.

Ivy has proposed the following technique for measuring the bleeding time: the cuff of a sphygmomanometer is placed on the arm and the pressure is raised to 40 mm. Hg. On an area of the forearm where there are no veins, the skin is pricked 2 to 3 mm. deep with a spring lancet. At 30-sec. intervals the blood is dried until it ceases to flow. The normal bleeding time is 2 to 6 min.

Prothrombin determination.[1] Quick had proposed a one-stage method for estimating whether there is sufficient prothrombin. Standard active thromboplastin and calcium are added to oxalated plasma; normal plasma clots in 11 to 13 sec. It is now known that retarded clotting observed with this method may be due not only to deficiency in prothrombin but also to insufficiency of accelerator factor or other causes. Two-stage methods are now used more frequently. In the first stage, plasma is defibrinated and thromboplastin and calcium are added; this is added in graded concentrations to a standard solution of fibrinogen and the concentration which clots in 15 sec. is noted. Several modifications have been proposed. Quick adsorbs prothrombin in plasma by means of tricalcium phosphate, and afterward it is separated by elusion in sodium citrate solution. This is added to rabbit plasma free of prothrombin together with thromboplastin and calcium. These methods are used in order to study the mechanism of coagulation in patients on whom operations have to be performed, especially in those suffering from hepatic diseases. They are also useful for following up the effects of dicumarol.

[1] WARNER, E. D., K. M. BRINKHOUSE, and H. P. SMITH, *Am. J. Physiol.*, **116**, 667, 1936; **125**, 296, 1939; *J. Exper. Med.*, **66**, 801, 1937; "Symposia on Blood Clotting and Allied Problems," Josiah Macy, Jr., Foundation, New York, 1948, 1949, 1950.

Other determinations. The platelet count and quantitative estimations of fibrinogen and Ca, or of the retraction of the clot, are also made in some cases.

HEMOSTASIS

Hemorrhage is stopped (hemostasis) principally by three mechanisms:

1. Local vasoconstriction, which can last from several minutes to several hours, is of great importance. Blood continues to circulate by collateral vessels that have not been damaged. The contracted vessels can resist considerable internal pressure without opening. When they again dilate, the damaged walls have been stopped up by other means (platelets, clots). Vasoconstriction is responsible for the stopping of hemorrhage in the rare cases of afibrinogenia (absence of fibrinogen) in which the blood does not clot. Vasoconstriction is produced by (*a*) direct response of the muscles in the blood vessels; (*b*) axonic and true reflexes; (*c*) vasoconstrictor substances, especially those released by the thrombocytes (serotonin).
2. Platelets "stick" to a damaged part of the vascular endothelium. If a small blood vessel is pricked with a needle, the opening is stopped up by a white thrombus formed by platelets agglutinated in a mesh of fibrin. Pathologic alterations in the endothelium also attract the platelets and cause intravascular clotting. A decrease in the platelet count is accompanied by soft, loosely adhering clots. This is observed, together with fragility of the capillaries, in the condition known as purpura hemorrhagica.
3. Blood clotting stops hemorrhage by occluding the opening in the damaged vessels. It is of primary importance, as is seen in patients who suffer from incoagulability of the blood. In these cases the slightest trauma causes severe bleeding—either large hematomas (bleeding in the tissues) or, if an open wound has been made, considerable loss of blood. The simplest surgical operations (*e.g.*, extraction of a tooth) can cause severe, and in some cases fatal, hemorrhage.

Blood clotting can be accelerated by local application of substances impregnated with thrombin or thromboplastin, but the latter is

less efficacious than the former. Substances used as a support for thrombin are gauze, fibrin in meshes or membranes, gelatine sponges, oxidized cellulose, etc. Thrombin and thromboplastin should not be injected intravenously, because this procedure may provoke serious accidents.

PATHOLOGIC ALTERATIONS OF BLOOD CLOTTING

Thrombosis. Intravascular clotting is called thrombosis. It is most frequently observed as a complication of phlebitis in the lower limbs after childbirth and surgical operations. Thrombosis is favored by (*a*) lesions in the endothelium of the blood vessels (phlebitis, arteriosclerosis, etc.) or the cardiac valves; (*b*) agglutination of the thrombocytes; (*c*) slow circulation due to several circulatory disturbances—increased viscosity of the blood, immobilization of the limbs, etc.; (*d*) metabolic disturbances (cachexia, obesity); (*e*) infections and allergy; (*f*) excess heat or cold, radiations; (*g*) coagulating agents (metals, snake venoms, etc.). Clots can be released from a thrombus and, passing through the right cavities of the heart, cause pulmonary embolism with more or less serious, and sometimes fatal, results. Heparin and dicumarol have been used with good results in the prevention of thrombosis after operations in which it is likely to occur, and in cases of myocardial infarction in order to prevent extension of coronary thrombosis.

Purpura. Spontaneous hemorrhage in the skin and mucosae is called purpura. It is due in some cases to fragility of the capillary walls. Thrombopenia is observed in some forms of purpura; bleeding time is prolonged, and the clot is soft and does not retract.

Hemophilia. This is a hereditary disease transmitted by females, who do not suffer from it themselves, to some of their male offspring according to the mendelian laws of heredity. Blood drawn from a vein of one of these subjects clots very slowly, sometimes only after several hours. Even small wounds cause severe hemorrhage; this tendency to bleed profusely lasts throughout life. The concentration of fibrinogen, calcium, thrombocytes, thromboplastin accelerator, prothrombin, heparin, and plasma proteins is normal. The blood clots if thrombin is added. Blood does not coagulate in these patients because thrombin is not formed owing to a deficiency in thromboplastic activity.

The following hypotheses have been advanced to explain hemophilia:

1. The platelets are abnormally resistant and do not set free thromboplastinogenase. This hypothesis has not been confirmed.
2. There is a congenital deficiency in thromboplastinogen (Quick).
3. There is an inhibitory factor which prevents the formation of thrombin.
4. Normal plasma contains a substance which shortens the coagulation time of hemophilic blood. This substance is not found in the plasma of subjects with hemophilia. The substance, together with the globulins, can be obtained from normal plasma.

When it is necessary to staunch hemorrhage in hemophiliac patients, the coagulation time can be shortened by (*a*) transfusion of normal blood (Weil, 1905); (*b*) injection of the antihemophilic factor in normal plasma; (*c*) local application of thrombin or thromboplastin.

Hepatic insufficiency. Removal of the liver is followed by a gradual increase in the coagulation time. Fibrinogen, prothrombin, and the thromboplastin accelerator diminish. Similar changes, especially the decrease in prothrombin and thromboplastin accelerator, have been observed in patients suffering from diseases of the liver. The liver continually replaces plasma proteins, fibrinogen, prothrombin, etc., which become deficient if the liver ceases to produce them in normal amounts.

Diseases with disturbances of blood clotting. In many cases there may be abnormal blood clotting, *e.g.*, in diseases which cause hypoproteinemia. Hypofibrinogenia or afibrinogenia, thrombocytopenia, deficiency of thromboplastin cofactor, increase in heparin or antithrombin may be the cause of disturbances in blood clotting.

BIBLIOGRAPHY

Allen, E. V., E. A. Hines, W. F. Kvale, and W. W. Barker, *Ann. Int. Med.*, **27**, 371, 1947.
Astrup, T., *Advances in Enzymology*, **10**, 1, 1950.
Best, C. H., *Harvey Lect.*, **36**, 66, 1941 (heparin and thrombosis).
Brinkhouse, K. M., *Medicine*, **19**, 329, 1940 (prothrombin).
Bordet, J., *Ann. Inst. Pasteur*, **34**, 561, 1920.
Chargaff, E., *Advances in Enzymology*, **5**, 31, 1945.
Eagle, H., *Medicine*, **16**, 95, 1937.
Ferguson, J. H., *Physiol. Rev.*, **16**, 640, 1936 (calcium in blood clotting).

Honorato, R., *Am. J. Physiol.*, **150**, 380 and 405, 1947 (thromboplastin cofactor).

Houssay, B. A., and A. Sordelli, *Rev. Inst. Bact. Dep. Nac. Hig.*, **1**, 485, 1918; **2**, 151, 1919; *J. de physiol. et de path. gen.*, **18**, 701, 1919 (snake venom and blood clotting).

Howell, W. H., *Physiol. Rev.*, **15**, 435, 1935.

Howell, W. H., *J. A. M. A.*, **117**, 1059, 1941.

Jorpes, J. H., "Heparin," Oxford, London, 1946.

Jorpes, J. H., "Symposia on Blood Clotting and Allied Problems," Josiah Macy, Jr., Foundation, New York, 1948, 1949, 1950.

Mann, F. D., E. S. Shonyo, and F. C. Mann, *Am. J. Physiol.*, **164**, 111, 1951 (hepatectomy and blood clotting).

Mann, F. D., and M. Hum, *Am. J. Physiol.*, **164**, 105, 1951 (thromboplastin cofactor).

Nolf, P., *Medicine*, **17**, 381, 1938.

Owren, P. A. "The Coagulation of the Blood," Oslo, 1947.

Owren, P. A., International Congress of Clinical Pathology, London, 1951.

Quick, A. J., "The Physiology and Pathology of Hemostasis," Lea & Febiger, Philadelphia, 1951.

Seegers, W. H., *Circulation*, **1**, 2, 1950 (prothrombin and AC globulin).

Seegers, W. H., *Pharmacol. Rev.*, **3**, 278, 1951 (drugs and blood clotting).

Silberberg, M., *Physiol. Rev.*, **18**, 197, 1938 (thrombosis).

Smith, H. P., and J. E. Flynn, *Ann. Rev. Physiol.*, **10**, 417, 1948.

Taylor, E. H., R. G. Minot, J. H. Lewis, *et al.*, *J. Clin. Investigation*, **24**, 698, 1945; **25**, 870 and 876, 1946. *Ann. Int. Med.*, **26**, 363, 1947 (antihemophilic globulin).

Wilander, O., *Skandinav. Arch. f. Physiol.*, Suppl. 15, 1939 (heparin).

Wöhlisch, E., *Ergebn. d. Physiol.*, **28**, 443, 1929; **43**, 174, 1940.

Zunz, E., in G. H. Roger and L. Binet, "Traité de physiologie normale et pathologique," vol. 7, Masson et Cie, Paris, 1934, p. 189.

Vitamin K and Prothrombin

Almquist, H. J., *Physiol. Rev.*, **21**, 194, 1941 (vitamin K).

Brinkhous, K. M., *Medicine*, **19**, 329, 1940.

Butt, H. R., and A. M. Snell, "Vitamin K," Saunders, Philadelphia, 1941.

Riegel, B., *Ergebn. d. Physiol.*, **43**, 133, 1940 (vitamin K)

Tanturi, C. A., and R. F. Banfi, *An. farm. bioquím. (Buenos Aires)*, **11**, 83, 1940; *Bol. d. Inst. clín. quir.*, **18**, 29 and 32, 1942; *Semana med.*, **49**, 736, 1942.

CHAPTER 9

Immunity

Definition. After death the body is invaded by the microorganisms living in its cavities; the activity of these organisms is the cause of postmortem putrefaction. The living organism, however, resists microorganisms (viruses, bacteria, fungi, and protozoa) which form the flora and fauna of its internal cavities or which invade it from the environment; this resistance is called immunity. The defense mechanisms against the soluble products of the activity of these microorganisms, animal and vegetable toxins, and foreign proteins are similar to those which protect the body from the invasion of microorganisms.[1]

Immunity is, therefore, a mechanism of defense against microorganisms, the soluble products of microorganisms, and foreign proteins. Immunity is also the response of the organism to agents that attack its integrity; it is therefore a special form of irritability. Immunity, moreover, can be considered as a metabolic or a digestive process, as it consists in the disintegration of substances that tend to alter the specific chemical composition of the body.

In a struggle between two living organisms, immunity represents the response of the attacked organism to the aggressor. If microorganisms are the attackers, the following may occur:

1. The microorganisms may succumb and be digested or eliminated without causing disease. These innocuous microorganisms are called saprophytes.
2. The microorganisms may cause disease. They are then called pathogens, and their capacity to harm the organism and resist its defense is known as their virulence.
3. Pathologic microorganisms can live in an organism without producing any apparent disease. Subjects which harbor these microorganisms are called "carriers."

[1] Only a brief summary of the principal facts of immunity will be made in this chapter. More detailed knowledge can be obtained in treatises on immunology.

An infection (invasion by microorganisms) may remain localized or may spread to the whole body (general infection). In many cases a microorganism has a tendency to localize itself in a determined organ or tissue, and by whatever way it enters the organism it goes to the preferred organ. This is known as the organotropism of the microorganism.

Natural and acquired immunity. An organism may be refractory to a given microorganism or toxic protein, in which case it will resist the foreign agent without showing any abnormality. If this characteristic is congenital or hereditary, it is called natural immunity. If it is the consequence of disease or of the accidental or provoked invasion by a microorganism, it is called acquired immunity.

Immunity can be acquired actively and passively. Active immunity is the result of

1. A successfully resisted attack by the same microorganism; *e.g.*, after recovery from typhoid fever there is a certain degree of immunity or increased resistance to the typhoid bacillus.
2. A successfully resisted attack by a closely related microorganism; *e.g.*, Jenner's vaccine or cow-pox provokes a mild disease (vaccine pustule) which immunizes the organism against smallpox.
3. Treatment with a living microorganism, the virulence of which has been reduced by drying (antirabies vaccine), heating (antianthrax vaccine), or chemical agents; *e.g.*, culture in a medium with bile diminishes the virulence of the bacilli of tuberculosis, which can then be used as a vaccine for the prevention of tuberculosis (B.C.G. vaccine).

4. Treatment with dead or dissolved microorganisms, *e.g.*, antityphoid vaccine.
5. Treatment with modified toxins; *e.g.*, vaccination against diphtheria is practiced by injecting anatoxin, a diphtheria toxin modified by formaldehyde.

Passive immunity is acquired by transmission from an animal with active immunity. Thus repeated injections of diphtheria toxin into a horse actively immunize the animal and provoke the formation of antitoxin against diphtheria in the plasma; the serum is called antidiphtheric serum. Injection of this serum into a man increases his resistance against diphtheria (passive immunity) for a short time, *i.e.*, until the antitoxin, which is a foreign substance, is destroyed. Antidiphtheric serum is used in the treatment of diphtheria to neutralize the toxin produced by the bacteria; it has little value in prevention of the disease.

Cellular immunity. Cells have, or can acquire, resistance to microorganisms and their products. Cells play an important part in immunity, either through phagocytosis or by the secretion of defense products. Phagocytic activity is especially developed in the leukocytes and the cells of the reticuloendothelial system; they first capture microorganisms that have invaded the organism and then dissolve them. Phagocytosis is the normal process of digestion in protozoa and myxomycetes; it is also observed in the cells lining the digestive tract of some invertebrates and in the migrating cells of many animals. Metschnikoff (1883) first studied phagocytosis and gave it its name. Phagocytosis plays a part in (*a*) the nutrition of certain invertebrates; (*b*) the reabsorption of embryonic and pathologic tissue or cellular remains; (*c*) immunity; (*d*) inflammation. Phagocytosis is favored by opsonins and bacteriotropins in the plasma, which act on bacteria and facilitate their absorption by the phagocytes.

Microorganisms produce dilatation of the blood capillaries and increase the permeability of capillary endothelium, so that plasmatic fluid passes out from the blood to the tissues, which become swollen (edema). Microorganisms also produce soluble substances which attract leukocytes and migrating histiocytes. This process is called inflammation. It is classically described by its four cardinal signs: "rubor, calor, dolor, and tumor"—signs due to vasodilatation and to the accumulation of fluid and cells. The living and dead leukocytes gathered in the inflamed focus form the pus found in abscesses or in the exudation of inflamed mucosae.

Menkin has reported the appearance of the following factors in inflamed tissues:

1. *Leukotaxin*, a substance that increases capillary permeability and attracts and stimulates migration of leukocytes.
2. *Necrosin*, a highly toxic euglobulin that kills the cells locally and damages the liver and kidneys if it passes into the circulation.
3. *Pyrexin*, a factor that provokes fever when introduced into the circulation.
4. *LPF, or leukocyte-promoting factor*, which arises in inflamed tissues, passes into the circulation, and releases granulocytes from the bone marrow.
5. *The leukopenic factor* which arises locally in the course of acute inflammation.

Some corticoids (desoxycorticosterone) and hormones (hypophyseal somatotrophin) stimulate the proliferative capacity of connective tissue. Other adrenal hormones (cortisone and hydrocortisone) have the opposite effect; they inhibit the building up of barriers of granulomatous tissue against invaders; they depress inflammatory reactions and open the way for the spread of infection (Selye).

HUMORAL IMMUNITY

Antigens and antibodies. Chemical substances produced by microorganisms provoke the appearance in the invaded organism of other substances which neutralize bacterial products. These reacting substances are called antibodies, and antigen is the name given to a foreign substance that provokes the formation of antibodies. Antigens are proteins of another animal or vegetable species.[1] Some proteins have little antigenic activity and seldom provoke the formation of antibodies; *e.g.*, insulin, protamine. Bacteria have several antigens (mosaic of antigens), some of them specific to one microorganism, others common to several species. Many toxic or nontoxic proteins of vegetable (abrine, ricin) and animal (venoms of snakes, spiders, and scorpions, serum proteins, etc.) origin have

[1] The lens of the eye is the only tissue capable of provoking the formation of antibodies in individuals of the same species.

antigenic activity and provoke the formation of antibodies.

Haptenes. Certain substances are "partial antigens"; they can unite with antibodies but do not have the capacity to stimulate the production of antibodies. They are more or less complicated carbohydrates or fats, or relatively simple substances. Combined with a protein, they acquire antigenic properties and provoke the formation of antibodies. Landsteiner has called them "haptenes." The pneumococcus, for example, owes its specific pathogenic properties to certain carbohydrates that do not produce antibodies. If these carbohydrates are joined to an antigenic protein, the new compound is capable of stimulating the formation of antibodies that neutralize the pathogenic effect of the carbohydrate.[1] Antibodies for nonantigenic substances are now prepared by uniting these to antigenic proteins.

Antitoxins. Bacteria excrete soluble toxins (exotoxins); they also have toxins (endotoxins) that remain within the microorganism. Repeated injections of increasing doses of a toxin provoke the formation of an antitoxin (Behring and Kitasato, 1890) which neutralizes specifically the toxin injected. Antitoxin is found in the pseudoglobulins of plasma or serum. These antitoxins are used in the treatment of diphtheria, tetanus, and the bites of snakes, spiders, or scorpions. Concentrated antitoxin is the pseudoglobulin isolated from antitoxic serum, separated from other inactive proteins, so that in a smaller mass there is more active and less inactive substance. Pseudoglobulin can be further purified by partial digestion with pepsin, because the active globulin is more resistant to digestion than the inactive globulin that accompanies it. This process gives very active serums which have little inert substance and which seldom provoke the accidents known as serum sickness. The γ globulin of the plasma or serum contains antibodies against many pathogenic agents. Once it has been isolated it can be concentrated to 20 or 25 times its original concentration. This γ globulin has been used successfully in the prevention of measles. Diphtheria antitoxin is found in the β_2 and γ globulins.

Agglutinins. In the course of certain diseases agglutinins for the microorganism causing the disease appear in the patient's serum. These agglutinins can be demonstrated by placing in a test tube the patient's serum and the bacteria causing the disease. After a short incubation period, the bacteria conglomerate in clumps, which are not dispersed by moderate shaking. The appearance of specific agglutinins in the serum of a patient is used in the diagnosis of infection, *e.g.*, Widal's test for typhoid fever. Agglutinins are produced by injecting bacteria or erythrocytes into an animal (*e.g.*, a rabbit); these antibodies are highly specific, *i.e.*, they agglutinate only the bacteria or the erythrocytes of the type injected, and not others. Nevertheless some of the bacteria have an antigen mosaic in which some of the antigens are common to other species. In this case the immune serum also agglutinates other bacteria which have antigens in common with the species used for immunization, but this agglutinating activity is not so intense as for the specific microorganism.

Precipitins. Repeated injections of a soluble protein provoke the appearance of a precipitating activity against the protein in the injected animal's serum. The immune serum is placed in a test tube, and when the protein is added, if there is precipitin in the serum, it becomes cloudy. These precipitins are highly specific against soluble bacterial protein (Kraus, 1897) and many nontoxic tissue proteins (Bordet, 1898). Precipitation tests are used in legal medicine to identify the nature of stains; thus from a blood stain there can be dissolved a protein that will give a precipitate with antihuman serum only if the stain was of human blood. This test has been used by Nuttall to establish the zoological relationship between different species, *e.g.*, between the domestic guinea pig (*Cavia domestica*) and the wild species (*C. cuis*).

Cytolysins. Repeated injections of animal cells into animals of another species will provoke the development, in the injected animal's plasma or serum, of the capacity to dissolve the injected cells. A typical example is that of hemolysins; *e.g.*, by injecting sheep erythrocytes into a rabbit, the serum of the injected animal acquires the power to dissolve (hemolyze) sheep's eryth-

[1] Many tissue extracts act as heterophil antigens; injected into a rabbit, they provoke the formation of hemolysins for sheep erythrocytes. Sordelli, Wernicke, Pico, Fischer, and Deulofeu (*Rev. Soc. argent. de biol.*, **1**, 186, 1920; **5**, 570, 1924) demonstrated that this antibody was active against lipids, but that lipids did not stimulate the formation of the antibody; the lipids joined to a protein had antigenic activity. This fundamental fact was later studied by Landsteiner, who developed his concept of partial antigens or haptenes.

rocytes. Cytolysins are made up of two components: one, which is specific for the cell and resists heat, is known as the amboceptor; the other, which is not specific and is destroyed by heating at 56°C., is called the complement or alexin. Bacteriolysins are cytolysins obtained by injecting bacteria.

ANAPHYLAXIS

Richet and Portier (1902) first described the phenomenon known as anaphylaxis. A toxic (Richet) or nontoxic protein, *e.g.*, horse serum, is injected into an animal; after an interval of one to three weeks the same protein is again injected with serious, sometimes fatal, results. The condition is called anaphylactic shock because of its suddenness and intensity. This hypersensitiveness is specific for the substance injected, which produces a state opposite to that of immunity or protection; hence the name of anaphylaxis used by Richet.

Anaphylactic shock. Anaphylactic shock has been studied in many species, but mostly in the dog (Richet), the rabbit (Arthus), and the guinea pig (Theobald Smith, 1906). Three steps must be considered: (*a*) the sensitizing dose; (*b*) the latent period, or incubation time; (*c*) the shock dose. The sensitizing substances are antigens, *i.e.*, proteins or substances combined with proteins (anaphylactogenic antigens). Very small doses are sometimes sufficient to produce sensitization.

The incubation time is usually two or three weeks. Hypersensitiveness lasts for several months or even years. If the animal is sensitized, the shock dose (preferably injected intravenously) provokes anaphylactic shock. Sensitization is strictly specific; the shock is provoked only by the substance to which the animal has been sensitized. Specificity is not so strict in the rabbit as in the guinea pig. Anaphylaxis can be used to identify the nature of a protein; thus three different proteins with specific sensitizing capacity have been found in horse serum. The shock dose produces the most intense reaction when injected intravenously, but it also provokes shock, although in decreasing order of intensity, when injected into the brain, peritoneum, or trachea, or subcutaneously.

Anaphylactic shock provokes reactions which vary according to the species but which are always the same whatever the antigen used. They are therefore conditioned by the species, not by the antigen. All the tissues react, but the signs observed are due to the contraction of plain-muscle fibers and to increased permeability of the endothelia. In the guinea pig acute anaphylactic shock causes death by asphyxia. The muscles in the bronchioles are contracted so that air goes into the lung but does not come out; on opening the thorax, the lungs remain distended. In the dog there is a typical triad: hypotension, leukopenia, and incoagulability of the blood. Hypotension is due to contraction of the suprahepatic veins, which causes stasis in the portal system. Leukopenia is due to the accumulation of leukocytes in the small blood vessels of the abdomen; a leukocyte count in peripheral and abdominal blood shows a marked difference, and abdominal vasoconstriction is followed by an increase in leukocytes in peripheral blood. Incoagulability is due to the secretion of heparin by the liver which forms antithrombin in the blood. The liver also releases into the circulation histamine; bradycinin, a slow-acting vasoconstrictor (Rocha e Silva); and fibrinolysin. Anaphylactic shock in the dog is similar to shock provoked by histamine, peptone, and other agents. In the rabbit anaphylactic shock causes hypotension due to constriction of the pulmonary blood vessels.

Passive anaphylaxis. If the serum from a sensitized animal is injected into a normal animal, the latter after a few hours becomes sensitized. An injection of the antigen provokes anaphylactic shock, which may be fatal. This experiment demonstrates the existence of a circulating antibody which can be transmitted to another animal. The latent period is necessary for the passive sensitization of the animal's cells, so they can react to the antigen, which only then will provoke shock.

Cellular anaphylaxis. Cellular sensitization can be demonstrated as follows: a sensitized dog is bled, and normal blood is transfused until the animal's blood has been almost completely replaced; a shock dose of the antigen is then injected and shows the animal is still sensitized (Manwaring). An even clearer demonstration is obtained by taking out the uterus of a sensitized guinea pig, washing it free from blood, and submerging it in Tyrode's fluid at 37°C. The tissue survives in good condition and responds to many drugs. When minute quantities of the

antigen to which it has been sensitized are added to the fluid, violent contractions are provoked in the isolated uterus. Other antigens do not produce this effect (Dale).

Desensitization. A very small dose of antigen does not provoke symptoms in a sensitized animal, and protects it from the effects of a larger dose injected a little later, which would otherwise have produced serious and perhaps fatal results. Desensitization can also be obtained by a very slow injection of dilute antigen. In these cases the organism reacts but not with the violent symptoms of shock. Besredka has advocated this method of desensitization in cases in which serum treatment must be given to patients who have been injected with the same serum some time previously.

Theories of anaphylaxis. Many theories have been proposed to explain anaphylaxis, but it is still unknown whether shock is due to flocculation of colloids, or an intense antigen-antibody reaction within the cells, or to the liberation of active substances. Intravenous injection of colloid solutions sometimes provokes shock very similar to anaphylactic shock, without any previous sensitization. During anaphylactic shock, histamine is liberated in the lung of the guinea pig and the liver of the dog. Histamine injection produces the same symptoms as anaphylaxis, but there is no definite proof that anaphylactic shock is due to the liberation of histamine.

ALLERGY

Allergy is an abnormal reaction of the organism to a foreign substance. Its most interesting manifestation is hypersensitiveness. Many people are hypersensitive to animal or vegetable allergens which when given by mouth, inhalation, or injection are capable of producing reactions. Typical among allergic phenomena are vasodilatation in restricted areas of the mucosae, increased capillary permeability with local edema or hemorrhage and contraction of plain muscles, producing hay fever, asthma, gastrointestinal disturbances, conjunctivitis, etc. Serum from these patients injected into other subjects or animals does not passively sensitize them to the allergen, and they do not react to it with the typical allergic syndrome.[1] Local sensitization can nevertheless be produced passively by intradermal injection into a normal subject of a small amount of serum (0.1 cc.) from a sensitized subject. Injection of the allergen into the sensitized area provokes an intense local reaction (itching, redness, swelling, edema); a neighboring area of normal skin does not react to the allergen in this way. This passive transmission of hypersensitiveness to a local area (Prausnitz-Küstner reaction) is due to the existence in the serum of special antibodies (reagins) which have a tendency to remain fixed where they have been injected. Allergic idiosyncrasy is frequently observed. Actual hypersensitiveness is not inherited, but there is a hereditary tendency to acquire hypersensitiveness and to react by an allergic syndrome.

Allergic reactions, in some cases, are more or less inhibited by antihistamine drugs, cortisone, and adrenocorticotrophin.

Bacterial allergy. Bacteria and other parasites provoke hypersensitiveness in the host to some of their constituent substances or excretions. Thus subjects infected with the tubercle bacillus react in a peculiar way if tuberculin is placed on the scarified skin (von Pirquet's reaction) or injected intradermally (Mantoux reaction); there is reddening, swelling, and itching, which does not occur in nonallergic subjects. Patients with a hydatid cyst react in the same way when hydatid fluid is applied to the skin by scarification or intradermal injection (Casoni's reaction).

BIBLIOGRAPHY

BORDET, J., "Traité de l'immunité dans les maladies infectieuses," Masson et Cie, Paris, 1939.

BOYD, W. C., "Fundamentals of Immunology," Interscience Publishers, Inc., New York, 1943.

MENKIN, V., Modern Concepts of Inflammation, *Science*, **105**, 538, 1947.

ROCHA E SILVA, M., "Histamina e Anafilaxia," Edigraf, São Paulo, 1946.

ZINSSER, H., *et al.*, "Immunity," The Macmillan Company, New York, 1939.

[1] Passive transmission of asthma has been observed in a few cases by transfusion of blood from an asthmatic subject to a normal one (Ramírez, Frugoni).

SECTION TWO

The Circulation

CHAPTER 10

The Circulation of the Blood

THE PROCESS OF cellular nutrition demands a continuous supply of nutritive substances and the subsequent removal of resulting waste products. In unicellular organisms and the simpler multicellular ones, this continuous exchange takes place directly between the cells and the environment. The currents due to differences in temperature levels or other causes, or the simple diffusion of substances in the fluid mass, or the displacement and movements of the organism if they can take place, are enough to ensure the renewal of the environment in the immediate neighborhood of the organism, and therefore the arrival of food and the removal of waste products.

The development of more complex organisms has been made possible by the simultaneous development of a vascular system that serves to carry to all the tissues a fluid, *i.e.*, the blood, the constant circulation of which assures to all the cells the transportation of material necessary for their activity. In vertebrates—the animals in which we are here most directly interested—the vascular system is made up of completely closed tubes, which in no place permit their contents, the blood, to come in direct contact with the cells. The system of tubes within which the blood circulates has different types of structures in its various sections. Three main types can be distinguished: arteries, capillaries, and veins. An important part of the vascular system is the mechanism of propulsion that moves the blood and keeps it circulating, in order that it may arrive at all the tissues in adequate amounts and at an appropriate pressure. The heart is the most outstanding feature of this propulsive mechanism, and for this reason it is considered the central organ of the circulation.

The vascular system, therefore, consists of a series of closed tubes, the arteries, capillaries, and veins, and a central organ of propulsion, the heart. The blood is driven by the heart into the arteries, from which it passes into the capillaries, then into the veins, which return it to the heart. The blood simply passes through the arteries and veins, and as it runs through these vessels there is no significant change in its composition or in its properties. On the other hand, as the blood passes through the capillaries there is an interchange of substances between the blood and the tissue fluids, and the chemical composition and the physical properties of the blood change considerably.

The circulation is an important factor in functional integration and equalization of the chemical and physical properties of the whole organism.

Arteriovenous anastomoses. In certain parts of the body such as the skin, the erectile organs, the balls of the fingers, etc., some of the arterioles open directly into the veins, without an intermediary capillary. This is a normal, but exceptional, arrangement of the blood vessels. Arteriovenous anastomoses play an important part in regulating the volume of blood which circulates through a given region.[1]

Aortic and Pulmonary Circuits. In mammals and in birds there are two complete circuits which begin and end in the heart. Venous blood arrives at the right auricle, passes into the right ventricle, and from there flows through the pulmonary artery to the lungs, where in the capillaries it undergoes the process of arterialization. This arterial blood flows through the pulmonary veins into the left auricle, passes into

[1] CLARK, E. R., *Physiol. Rev.*, **18**, 229, 1938.

the left ventricle, and is sent through the aorta and its branches to the capillaries in the tissues and organs, whence it flows back through the veins into the right auricle. The pulmonary circuit, or lesser circuit, begins in the right ventricle and ends in the left auricle; the aortic circuit, or greater circuit, begins in the left ventricle and ends in the right auricle. The blood flows continuously and successively through these two circuits.

HISTORICAL REVIEW

The first conclusive demonstration of the nature and the principal mechanisms of the circulation of the blood was given by William Harvey in his classic book, *Exercitatio anatomica de motu cordis et sanguinis in animalibus*, published in 1628 in Frankfurt. Harvey developed his ideas by observations and experiments that were made with a faultless method and followed a remarkably logical plan. It is therefore not surprising that his book can be read after more than three centuries, not only with admiration, but also without having to change any of his main conclusions as a result of later discoveries. Harvey's book is one of the most important contributions to science of all times, although its length is only 72 pages.

Harvey's observations and experiments put an end, although not without difficulty, to the vague and inaccurate empirical ideas of his day. Several isolated facts of importance in respect to the circulation of the blood had been known before Harvey made his contribution. Vesalius (1543) described the course of the veins and the anatomy of the heart. Fabritius described the valves in the veins, but he was not aware of their functional significance. Michael Servet (1551) and later Mateo Realdo Colombo (1555) described the passage of blood from the right ventricle through the lungs into the left auricle and ventricle and had some idea of the process of hematosis.[1]

After Harvey, Malpighi (1661) discovered the capillaries and another important stage was completed. Harvey had stated that the blood passed from the arteries to the veins, but he had not said how this was accomplished. The discovery of the capillaries, made possible by the development of the microscope, filled this gap in the knowledge of the circulatory system.

Further historical data will be given in the appropriate places in the following chapters.

[1] According to recent historical research (Haddad and Khairallah, *Ann. Surg.*, **104**, 1, 1936), in the twelfth century Ibn Naffis, an Arab physician, had stated that the blood went from the right ventricle to the left auricle through the lungs.

GENERAL LAWS OF THE CIRCULATION

The general laws of the circulation of the blood will be stated briefly here; later in the respective chapters they will be considered in detail.

The law of pressure. The pressure exerted by the blood on the walls of the blood vessels is determined by the cardiac output in unit time and by the peripheral resistance to its circulation. It is highest above the section that exercises the greatest resistance, *i.e.*, in the arteries. It drops suddenly at the level of the capillaries and continues to fall gradually along the veins, reaching a minimum in the auricles. This pressure gradient is the reason why the blood passes from the ventricles to the auricles in the general and pulmonary circuits.

The law of velocity. The velocity at which the blood circulates is dependent on the caliber of the vascular bed—not on the diameter of each blood vessel, but on that of all the divisions of the main trunk at a given distance from the heart. The ramification of the vessels is such that the sum of the cross-sectional areas of the branches is greater than the cross-sectional area of the parent vessel. Therefore the vascular bed increases as the distance from the heart is greater, reaching a maximum at the level of the capillaries. On the venous side it is progressively reduced as the distance to the heart diminishes, and it is at a minimum at the openings in the auricles. The vascular bed, in both the general and the pulmonary circuits, could therefore be represented by two truncated cones joined by their bases. One of the cones would represent the arterial side; the other, the venous side; and the area where they join, the capillaries. Once this idea of the vascular bed is clear, and taking into account that the velocity of a current in any stream is inversely proportional to the cross section of its bed, it is easy to understand the following law of velocity in the circulatory system: *the velocity of the blood diminishes in the arteries as the distance from the heart increases, it is at a minimum in the capillaries, and it increases progressively on the venous side as it comes nearer to the heart.* The total cross-sectional area of the veins where they end in the auricles is greater than that of the arteries which emerge from the ventricles; therefore the velocity of the blood flow is greater in the arteries than in the veins.

The law of volume flow. The amount of blood passing through a cross section of the circulatory system in unit time is called the volume flow. The circulation is normally maintained without undue stagnation of blood at one point and insufficient filling at others when the following law of volume flow is fulfilled; the amount of blood passing in a given time through a complete cross section of the circulatory system is the same as that passing through a complete cross section at any other point.

If 5 liters of blood passes through the aorta, then at the same time an equivalent amount will pass through the pulmonary artery and the aggregate of the systemic capillaries, etc. The small changes produced by the active modifications in the caliber of the vessels in a limited section are compensated by adequate reactions in other sections.

BIBLIOGRAPHY

BETHE, A., G. VON BERGMANN, G. EMBDEN, and P. ELLINGER, "Handbuch der normalen und pathologischen Physiologie," 7, Springer, Berlin, 1926.

FOSTER, M., "Lectures on the History of Physiology," Cambridge, New York, 1924.

HARVEY, W., "The Anatomical Exercises of William Harvey: De motu cordis, 1628; De circulatione sanguinis, 1649: The First English Text of 1653 Now Newly Edited by Geoffrey Keynes," Nonesuch, London, 1928.

IZQUIERDO, J. J., "Harvey, Iniciador del Método Experimental," Ciencia, Mexico, 1936.

LUCIANI, L., "Fisiología dell'uomo," Società Editrice Libraria, Milan, 1923.

ROGER, G. H., and L. BINET, "Traité de physiologie normale et pathologique," vol. 6, Masson et Cie Paris, 1932.

TIGERSTEDT, R., "Physiologie des Kreislaufes," Walter de Gruyter & Company, Berlin and Leipzig, 1921.

WILLIUS, F., and T. J. DRY, "A History of the Heart and the Circulation," Saunders, Philadelphia, 1948.

CHAPTER 11

The Physiologic Properties of the Heart

The heart is responsible for the force necessary to make the blood circulate. It is a hollow organ with muscular walls and has a system of valves so disposed that when the heart muscle contracts the blood that fills the cardiac cavities is driven forward from the venous to the arterial side.

The fundamental properties of cardiac muscle fibers are of essential importance in the function of the heart.

THE STRUCTURE OF THE HEART

The heart has certain structural features that must be known in order to understand many aspects of its function. The heart walls are made up mainly of muscle fibers of a special type. These are cylindrical, with a central nucleus and there are protoplasmic bridges between neighboring fibers, so that the heart muscle as a whole constitutes a syncytium. Continuity between the auricular and ventricular muscles is established by a special bridge, the bundle of His.

Cardiac muscle fibers have the longitudinal striation common to all muscle fibers and cross striations similar to those of skeletal muscle. This similarity in the histologic aspect of cardiac and skeletal muscle fibers does not extend to their physiologic properties. In its contraction, its response to innervation, and the chemical aspects of its activity, cardiac muscle differs fundamentally from skeletal muscle.

The electron microscope shows that in each myocardial fiber there are 300 to 700 myofibrils; these are the smallest contractile units. Capillaries are seen to enter the myocardial fiber carrying blood to the sarcoplasma, nuclei, and myofibrils. Large numbers of sarcosomes (mitochondria and chondriosomes) are seen scattered among the myofibrils and near the poles of the nuclei, but never within the myofibrils.[1]

THE PROPERTIES OF THE CARDIAC MUSCLE FIBER

Four fundamental properties are usually studied in the cardiac muscle fiber: rhythmicity, conductivity, excitability, and contractility.

Rhythmicity. The cardiac fiber has the power of originating within itself the impulse that makes it contract. This is one of the first properties to appear in the course of development. As soon as the heart is formed in the embryo, before the fibers acquire their typical histologic aspect, it contracts rhythmically, and this rhythmic beat persists even after it has been separated from the organism.

After the development of the heart is complete, not all its parts have the same rhythmicity. If the heart is removed from the body it will continue to beat for some time. The ventricles are the first to stop beating, then the left auricle stops, and finally—sometimes after more than half an hour—the right auricle comes to a standstill. For this reason the right auricle has been called the *ultimum moriens.*

After the auricular contractions have ceased, careful observation will reveal, even more than an hour after the heart has been separated from the body, a small circumscribed area of activity at thejunction of the superior vena cava with the right auricle. In this area, as will be explained later, there is a small mass of myocardial

[1] Kisch, B., and J. Bardet, "Electron Microscopic Histology," Brooklyn Medical Press, New York, 1951: Kisch, B., *Exper. Med. & Surg.*, **9**, 333, 1951.

tissue that is a vestigial equivalent of the venous sinus of fishes and amphibians. In these animals the sinus is also the cardiac segment with the highest rhythmicity.

In the lower vertebrates the heart has greater rhythmicity than in mammals. The heart of a fish or an amphibian will continue to beat regularly for several minutes, and sometimes hours, after it has been separated from the body. The mammalian heart subjected to this treatment will beat only a few times with all its cavities contracting in a coordinated sequence.

The activity of the isolated heart can be prolonged, even in mammals, by means of special techniques that assure the circulation of nutrient fluids through the heart. Defibrinated blood, Locke's fluid, or other fluids (to be mentioned at the end of this chapter) are used for this purpose. A mammalian heart, including that of man, can be kept beating for several hours, and amphibians' hearts for several days.

Because of this property of rhythmicity, the heart is automatically and periodically stimulated into activity. This is sometimes called the "chronotropic" property of the heart (Greek *χρονος*, time, and *τρὸπειν*, to turn), referring thus to the fact that it marks the time or pace of the heartbeat.

Conductivity. The stimulus that activates the cardiac muscle has its origin in a circumscribed area, but it spreads throughout the heart muscle thanks to its conductivity—a property of the myocardial fiber itself, as was demonstrated by Engelmann.[1]

Conductivity is a property of all the heart muscle, but it is especially developed in the bundle of His and its branches and in Purkinje's tissue.

The excitatory state is transmitted along the cardiac muscle in a way similar to that by which a nerve impulse is propagated along a nerve fiber. There is conduction without decrement (see Chap. 68). If the vitality of the tissue is diminished in any place, whatever the cause (anoxia, intoxication, pressure, inflammation), the excitatory state is transmitted at a lower speed; if the abnormal process is sufficiently severe, conductivity will be completely abolished. Successive stimuli cannot spread through the heart unless there is a certain interval between them.

The speed of transmission of the excitatory

[1] ENGELMANN, T. W., *Arch. f. d. ges. Physiol.*, **11**, 466, 1875.

state is dependent on the heart's conductivity; for this reason this is sometimes known as the "dromotropic" (Greek *δρομος*, running) property of the heart.

Excitability. The heart responds not only to the stimulus originated by its own rhythmicity, but also to external stimuli of different kinds. The heart has therefore the property of excitability. Many extraneous agents can act as stimuli; *e.g.*, mechanical (friction, a blow, pinching); electrical (induced and direct currents, a condenser discharge); thermal (heat, cold); chemical; etc.

The heart is not equally excitable during all the phases of its activity. During systole (contraction) the heart becomes inexcitable for a certain period known as the absolute refractory period, but, as the state of contraction ceases, excitability returns. Excitability increases rapidly very early in diastole, reaching a maximum that is maintained until the next beat.

Maximum excitability, *i.e.*, the lowest threshold, is reached after a few oscillations during the phase of rising excitability (relative refractory period), but a phase of supranormal excitability is not observed at any moment[1] (Fig. 17).

In subjects with heart disease, very premature contractions of the ventricle, *i.e.*, extrasystoles with QRS complexes occurring before the T wave of the preceding contraction has ended (see Chap. 16), have been observed. This phenomenon has been interpreted as proof of the existence of a period of supranormal excitability.[2]

This continuous variation in excitability makes a quantitative estimation of this property extremely difficult, as it is not easy to determine it at exactly the same moment of different cardiac cycles, a necessary condition for making valid comparisons between several observations. This is particularly true when it is necessary to measure the effect on cardiac excitability of drugs and other agents capable of changing excitability.

Between the moment at which the stimulus acts and the response of the heart there is a brief interval known as the latent period. In the dog it is approximately 30 msec. at the moment of the cycle when excitability is at its highest, but

[1] ORÍAS, O., C. MCC. BROOKS, E. E. SUCKLING, J. L. GILBERT, and A. A. SIEBENS, *Am. J. Physiol.*, **163**, 272, 1950.

[2] SMIRK, F. H., *Brit. M. J.*, **11**, 23, 1949.

it is more prolonged during the relative refractory period, increasing as the stimulus is applied nearer to the absolute refractory period. Stimuli applied close to the end of the absolute refractory period may have a latency of nearly 150 msec. For this reason, however early in the cycle

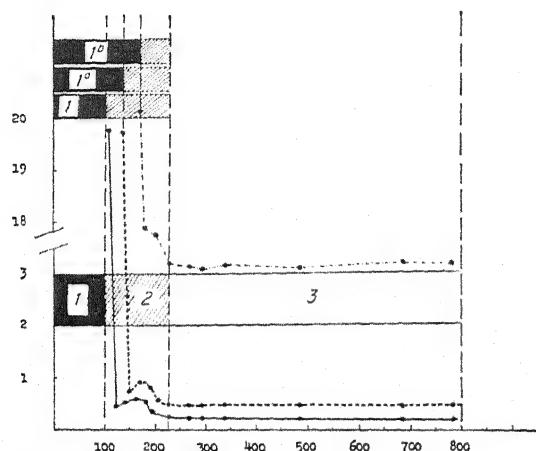



FIG. 17. Excitability of the ventricle in the dog. Abscissa: time, in milliseconds, between the stimulus provoking a beat and the test stimulus. Ordinate: intensity, in milliamperes, of test stimulus. Threshold for test: rectangular shocks of 0.1 msec. duration (dot-dash line), of 6 msec. (broken line), and of 13 msec. (solid line). 1, absolute refractory period; 2, relative refractory period; 3, phase of complete recovery of excitability. (*Orías, O., et al., Am. J. Physiol., vol. 163, p. 272, 1950.*)

a stimulus is applied, the response tends to occur at approximately the same moment, *i.e.*, toward the end of the relative refractory state, and there is no way of obtaining it earlier (Orías *et al.*). The relative refractory period is in fact *relative*, because a response can be obtained if a sufficiently strong stimulus is applied; in a certain sense, however, it is an *absolute* refractory period, because until it is ended there is no response.

Figure 17, designed with data taken from one of many experiments, shows that the myocardium is in the absolute refractory phase only during the first moments of systole, as can be seen if very strong stimuli are applied. By the time the heart enters into diastole, excitability has been almost completely recovered. Using stimuli of maximal strength, the absolute refractory period becomes longer if the duration of the stimulus is shortened (Fig. 17, 1, 1*a*, and 1*b*). The absolute refractory period is prolonged at the expense of the relative refractory period.

Changes in excitability cause variations in the threshold, *i.e.*, the minimum strength of stimulus needed to obtain a response. For this reason this is sometimes called the "bathmotropic" (Greek βαθμος, threshold) property of the heart.

Contractility. In response to the stimulus originated by its own rhythmicity, or to an external stimulus, the heart contracts.

Contractions due to the heart's rhythmicity are the spontaneous beats of the heart, and their succession is known as the cardiac rhythm.

External stimuli provoke a response only after the absolute refractory period has ended. The forced contraction so originated is called a premature beat or extrasystole.

Contractility is also known as the "inotropic" property of the heart (Greek ινος, fiber).

EFFECTS OF STIMULATING AGENTS ON THE MYOCARDIUM

Some special features of the response of the cardiac muscle fiber to different stimuli will be considered under this heading. The most convenient, and most frequently used, stimulus is an induction shock from an induction coil. The spontaneous heartbeat interferes with the experimental observation; therefore an amphibian heart is used, in which the beat is stopped by placing a tight ligature on the sinoauricular junction. Strips cut from mammalian hearts can also be used. All the phenomena described in the

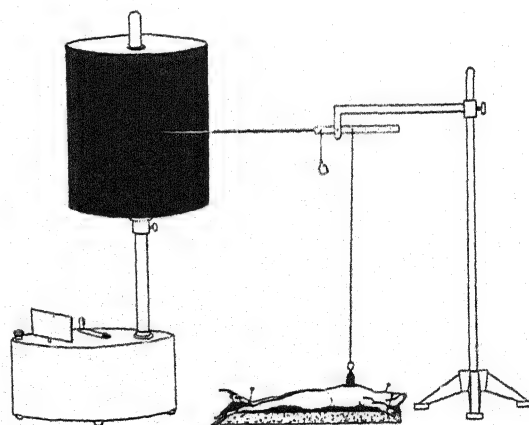



FIG. 18. Apparatus for obtaining a suspension cardiogram

following paragraphs can be demonstrated with a very simple apparatus (Fig. 18).

The "all-or-nothing" law. When a skeletal muscle is stimulated with a series of stimuli of increasing strength, the following phenomena are observed:

1. A very weak stimulus does not provoke a response.
2. On progressively increasing the strength of the stimulus, at a certain strength a very weak contraction will be observed. This is the minimum or threshold stimulus.
3. If the strength of the stimulus is further increased, the muscle responds with contractions of increasing strength, until a maximum is reached. Any further increase above this maximum stimulus does not provoke an increase in the strength of the contraction (Fig. 19).

The same experiment performed with cardiac muscle does not give exactly the same results. The stimulus, in this case, must also reach a minimum strength to provoke a response, but once the threshold has been attained, however much the strength of the stimulus is increased, the heart responds with contractions of constant strength (Fig. 19). The heart responds to the "all-or-nothing" law (Bowditch, 1871),[1] because when the stimulus has a threshold strength, the heart gives the maximum response.

Skeletal muscle
Myocardium
1 2 3 4 5 6

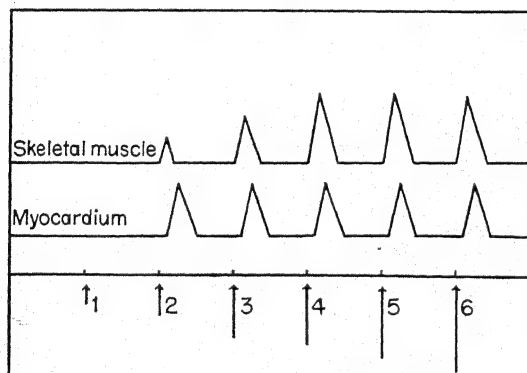



FIG. 19. The "all-or-nothing" law. The myocardium (cardiac muscle) responds with a maximal contraction, and the skeletal muscle with a submaximal contraction, to the threshold stimulus (2). To obtain a maximal contraction from the skeletal muscle the strength of the stimulus must be increased (4). With a subliminal stimulus neither the cardiac nor the skeletal muscles respond (1).

This particular form of response observed in cardiac muscle is due to the fact that all its fibers constitute a syncytium, and therefore the excitatory process is always propagated throughout the whole heart. In skeletal muscle, as the strength of the stimulus is increased, the excitatory state is propagated to a wider area and stimulates a larger number of fibers. The strength of the response depends on the number of fibers stimulated; therefore it will be greater as the stimulus is stronger. Individual muscle fibers also respond to the "all-or-nothing" law (see Chap. 66).

[1] BOWDITCH, H. P., *Ber. Sächs. Ges.*, **23**, 652, 1871.

1 2 3 4 5

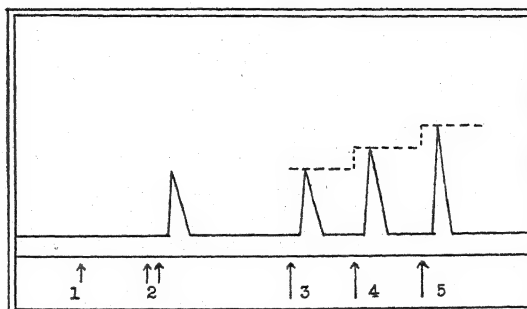



FIG. 20. Latent addition and "staircase phenomenon." Diagram of a suspension cardiogram. 1, one subliminal stimulus evokes no response; 2, two subliminal stimuli with an appropriate interval between them provoke a contraction (latent addition); 3, 4, and 5, stimulation with stimuli of constant strength at adequate intervals has produced increasingly greater responses ("staircase phenomenon").

Summation of stimuli. If a stimulus is below the threshold strength it will not provoke a response (subliminal stimulus), but if it is repeated two or more times at adequate intervals, a contraction may be observed. This is known as latent addition or summation of stimuli (Fig. 20). The same fact is observed in skeletal muscle and other tissues (see Chap. 66).

The staircase phenomenon. When a heart that has been at rest for some time is excited by a series of stimuli of threshold (or above-threshold) strength, repeated approximately every 10 sec., the three or four contractions first observed are of increasing magnitude. If horizontal lines at the level of the summit of each contraction curve are traced and joined by vertical lines, a staircase of three or four steps results (Fig. 20); hence the name of "staircase phenomenon" given to this observation.

The cause of this increasing response to a stimulus of constant strength is not known. It has been attributed to the accumulation of metabolic products resulting from the preceding contraction, which would place the muscle in condition to respond more vigorously.[1]

[1] DALE, A. S., *J. Physiol.*, **75**, 1, 1932.

If a longer interval (about 15 sec.) is allowed between the stimuli, the contractions are all of equal strength; there is no "staircase."

Circular excitation. Mines[1] observed some peculiar facts in muscular rings cut out from the ventricle of amphibians. These observations have been of great importance in the interpretation of certain functional anomalies of the cardiac contraction. Moderate pressure was applied on one point of the ring and the muscle was stimulated in the neighborhood (on either side) of the blocked area. A wave of contraction was thus initiated at the stimulated spot and transmitted toward the parts that were not blocked. After the wave had been started the block was removed. When the contraction wave, having traveled around the ring, arrived at the previously blocked area, this area was stimulated by the wave and transmitted the excitatory state to parts of the ring which had already been contracted. In this way the wave continued to travel around the ring for a relatively long time. This particular type of excitation and contraction is known as a "circus movement."

FIG. 21. Circus movement. A wave of excitation (black) passes through a ring of muscle, leaving behind it a refractory state (stippled). The arrow marks the direction in which the wave spreads. (*After T. Lewis.*)

The cause of this movement is the following: The blocked area prevents the spread of excitation in one direction, so the excitatory state can be transmitted only in the other direction. If the block is removed after the wave of contraction has progressed a short way, it cannot be excited by a "backward" spread of excitation because it is protected by an area that is still in a partially refractory state as a result of the passage of the wave. To stimulate the previously blocked area, the wave must travel around the ring. By that time the refractory state of the areas that have already contracted will have ended, so they will be in condition to respond to the excitation transmitted through the previously blocked area. The wave of contraction thus travels around the ring several times because it finds tissue in an excitable state before it and leaves tissue in a refractory state behind it (Fig. 21).

[1] MINES, G. R., *J. Physiol.*, **46**, 349, 1913.

Mayer[1] had made a similar observation in muscular rings taken from jellyfishes.

Fibrillation. In certain conditions the heart muscle fibers do not contract as in a normal beat, *i.e.*, more or less simultaneously. The contraction of the fibers is not coordinated; therefore the whole muscle does not contract, but some fibers, or groups of fibers, are in contraction, and other fibers, or groups of fibers, in relaxation (Fig. 22). The heart muscle seems to tremble or to be in a state of continuous convulsive movement. If a large part of the heart is involved, *e.g.*, all the fibers of the ventricles in a dog, this form of activity, known as fibrillation, has a tendency to persist. If the cardiac mass is small (a fragment of ventricle, or a small heart) fibrillation disappears after a few seconds. In small fragments of heart (1 gm. or less), fibrillation cannot be provoked.[2]

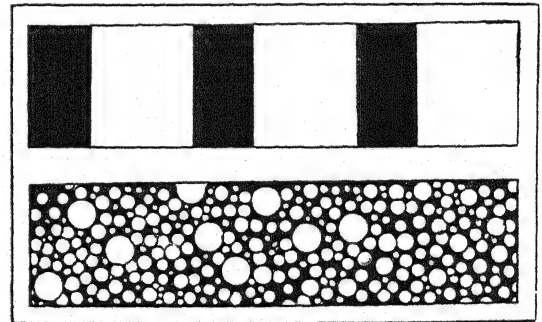



FIG. 22. Diagram illustrating normal contraction of the heart (above) and fibrillation (below). Above, the black bands represent the systole with all the fibers contracting simultaneously; the white bands represent the diastole with all the fibers relaxed. Below, there are fibers in systole (in black) and others in diastole (in white) at all times.

Fibrillation can be produced in several ways. The simplest is to stimulate the heart with a tetanizing induced current, or an alternating current of sufficient intensity. Isolated electrical stimuli in certain conditions can also provoke fibrillation, and the same result can be obtained

[1] MAYER, A. G., "Papers from Tortugas Laboratories," **1**, 115, Carnegie Institution, Washington, D.C., 1908.
[2] MARTÍNEZ, C., *Medicina, Buenos Aires*, **4**, 109, 1944.

with other stimuli: mechanical (pinching, blows), thermal (heat, cold), and chemical (caustic substances). The obstruction of a main branch of one of the coronary arteries is a frequent cause of fibrillation. Excessive doses of several drugs, such as digitalis, adrenaline, some of the anesthetics, and potassium, have the same effect.

There has been much discussion on the nature and mechanism of fibrillation, and as yet there is no unanimous agreement on the subject. Probably fibrillation is a form of cardiac activity caused by a particular condition in which the excitability and conductivity of the cardiac fibers are not homogeneous. Its initiation depends on the moment of the cardiac cycle at which the stimulus is applied. If the stimulus arrives when all the fibers are equally excitable and their conductivity is the same, they will contract simultaneously and a normal heartbeat will occur. On the contrary, if the stimulus is applied when only part of the fibers are excitable and can conduct the excitatory process, while others are in a refractory state, owing to the complex structure of the heart muscle, there will be a network of excitable fibers interwoven with a network of refractory fibers. Some of the fibers will then contract while others remain quiescent, *i.e.*, fibrillation will be provoked. The condition will have a tendency to persist, because there will always be some active fibers capable of transmitting the excitatory process to others that have just recovered from the refractory state.

Wiggers' work[1] is an important contribution toward solving the problem of the production and nature of fibrillation. He and his collaborators have shown that one isolated stimulus can provoke fibrillation if it acts at a particular moment of the cardiac cycle. This moment, appropriately called the "vulnerable phase" of cardiac activity, coincides with the phase of fluctuating excitability of the relative refractory period,[2] when some of the cardiac fibers are still refractory and others have more or less recovered their excitability.

Nutritional disturbances limited to certain areas of the heart produce a condition of heterogeneity or disparity between them and thus become factors that cause fibrillation. A similar effect is produced by several drugs and toxic substances, because they affect unequally the properties of the different fibers.

[1] WIGGERS, C. J., *Bol. Acad. nac. de med. de Buenos Aires*, 567, 1938; *Am. Heart J.*, **20**, 399, 1940.

[2] BROOKS, C. McC., O. ORÍAS, J. L. GILBERT, A. A. SIEBENS, B. F. HOFFMAN, and E. E. SUCKLING, *Am. J. Physiol.*, **164**, 301, 1951.

Prinzmetal and his associates[1] have applied cinema techniques to the study of fibrillation. If films obtained at a rate of 3,000 pictures per second are projected at a normal rate of 16 pictures per second, a very slow-motion effect is obtained (speed is reduced 180 times). Even thus fibrillation, according to Prinzmetal and his associates, appears as a completely disorganized activity, which proceeds without rhyme or reason, spreading to all the contractile elements of the auricle and different from all other normal or abnormal functional states of the heart. At no time did they observe anything remotely similar to circular excitation.

Flutter. In this form of activity the heart muscle shows a series of contractions, one following immediately after the other. Excitability and conductivity are in such a condition that the excitatory process, after having provoked the contraction of a fiber, can immediately stimulate the same fiber, because the refractory state has disappeared. It is a condition similar to the circus movement already described. Prinzmetal and his associates conclude, however, from their studies with the cinema technique mentioned earlier that auricular flutter is due to the periodic discharge of impulses of high frequency from a single focus, which spread simultaneously in all directions.[2] They deny, therefore, the existence of a circus movement.

THE TONUS OF THE MYOCARDIUM

It is difficult to determine whether cardiac muscle has true tonus. At any rate it does not have a tonus actively produced and completely dependent on nervous activity, comparable with the tonus of skeletal muscles. The innervation of the heart, provided by the autonomic nervous system, differs considerably from the somatic innervation of skeletal muscles. Moreover, in the diastole there are no electrical phenomena denoting activity, while even at rest there are electrical variations in the skeletal muscles, corresponding to their tonic activity.

[1] PRINZMETAL, M., E. CORDAY, I. C. BRILL, R. W. OBLATH, and H. E. KRUGER, "The Auricular Arrhythmias," Charles C Thomas, Springfield, Ill., 1952.

[2] *Ibsd.*

According to Starling[1] what is usually called the "tonus" of the heart is synonymous with its physiologic condition. A heart in good condition has a high tonus and empties itself almost completely at each systole even when it receives a large amount of blood during diastole. A heart with a low tonus is in a condition of fatigue. It is dilated at the end of systole and still contains a large amount of residual blood.

Table 15. Percentage Composition of Physiological Saline Solutions

	Ringer (frog)	*Locke*	*Tyrode*	*Dale*	*Lovatt-Evans**
Water	100.0	100.0	100.0	100.0	100.0
NaCl	0.65	0.9	0.8	0.9	0.85
KCl	0.014	0.042	0.02	0.042	0.042
$CaCl_2$	0.012	0.024	0.02	0.024	0.024
$NaHCO_3$	0.02	0.01–0.03	0.1	0.05	
NaH_2PO_4	0.001	....	0.005		
$MgCl_2$	....	....	0.01	0.0005	
Glucose	0.2	0.1–0.25	0.1	0.05	0.1
Na_2HPO_4	....	....	....	....	0.06

* 0.02–0.06 cc. of normal H_3PO_4 is added to each 100 cc.

According to Wiggers,[2] the tonus of the myocardium is the equivalent of its distensibility. On receiving the venous inflow at a given pressure, a heart with a high tonus is distended less, and therefore receives a smaller volume of blood, than one with a low tonus. The idea is here implicit that the condition of the cardiac fiber regulates its distensibility. It is nevertheless very difficult to demonstrate changes in the distensibility of a normal myocardium.

EQUILIBRATED SALT SOLUTIONS

Many years ago Ringer[3] first demonstrated the great importance of certain salts for the maintenance of rhythmicity and other properties of the heart muscle. Simple saline solutions cannot keep the heart beating for any length of time. A fragment of tortoise heart placed in 0.7 per cent NaCl solution begins to beat rhythmically, but soon the contractions weaken and finally they cease, the heart fragment remaining flaccid and completely relaxed. If now a very small quantity of $CaCl_2$ is added to the solution, the heart again begins to beat, but gradually relaxation is less complete after each contraction, until finally the fragment stops beating, remaining tonically contracted. If a very small amount of KCl is now added, rhythmic contractions again appear, and the fragment can continue to beat for several hours, although the solution does not contain any nutritive substance capable of providing energy for the contracting muscle. The mechanism by which the different ions act on the heart is still unknown.

Since these now classic experiments were reported, many other experiments have been made in which the isolated heart and other organs have been kept active and in a good condition. The proportions of Na^+, Ca^{++}, and K^+ vary according to the particular conditions of the organ studied. In the case of tissues or organs of homoiothermic animals, it is also necessary to keep them at a temperature of 38 to 39°C. and to oxygenate the perfusion fluid. Locke[1] showed that by adding glucose it was possible to keep the heart beating for a much longer time, and even to revive the activity of a heart which had ceased to beat in a saline solution free of glucose. Sodium lactate can replace glucose with advantage.

Besides the ionic equilibrium discussed above, a perfusion fluid must have an adequate osmotic pressure and acid-base equilibrium, and it must be at a temperature suitable for the tissue examined.

[1] Starling, E. H., "Principles of Human Physiology," 9th ed. (Lovatt-Evans), J. & A. Churchill, London, 1945, p. 553.

[2] Wiggers, C. J., "Physiology in Health and Disease," 4th ed., Lea & Febiger, Philadelphia, 1944.

[3] Ringer, S., *J. Physiol.*, **4**, 29, 1883.

BIBLIOGRAPHY

Clark, A. J., "Comparative Physiology of the Heart," Macmillan, New York, 1927.

De Boer, S., *Ergebn. d. Physiol.*, **21**, 1, 1923 (fibrillation).

[1] Locke, F. S., *J. Physiol.*, **18**, 333, 1895.

DEMOOR, J., "Ce que réprésente l'automatisme du coeur," Gaston Doin, Paris, 1923.

EYSTER, J. A. E., and W. J. MEEK, *Physiol. Rev.*, **1**, 1, 1921 (activation of the heart).

FRÉDÉRICQ, H., "Aspects actuels de la physiologie du myocarde," Presses Universitaires de France, Paris, 1927.

FRÉDÉRICQ, H., *Physiol. Rev.*, **8**, 501, 1928 (chronaxie).

GARREY, W., *Physiol. Rev.*, **4**, 215, 1924 (auricular fibrillation).

HABERLANDT, L., *Ergebn. d. Physiol.*, **25**, 86, 1926 (activation of the heart).

MEEK, W. J., *Physiol. Rev.*, **7**, 259, 1927 (heart tone).

TIGERSTEDT, R., "Physiologie des Kreislaufes," Walter de Gruyter & Company, Berlin and Leipzig, 1921.

WIGGERS, C. J., Dynamics of Ventricular Contraction under Abnormal Conditions, *Circulation*, **5**, 321, 1952.

WINTERBERG, H., in Bethe's "Handbuch der normalen und pathologischen Physiologie," **7**, 663, Springer, Berlin, 1926 (fibrillation and flutter).

CHAPTER 12

The Shape, Consistency, Movements, and Volume of the Heart

THE HEART IS a hollow organ, with elastic walls of unequal thickness, and with compartments, the contents of which vary in amount from one moment to another.

The beating heart can be observed directly in poikilothermic and homoiothermic (see Chap. 48) animals. The former have the advantage of being easier to handle, and their hearts can be exposed with a very simple technique. Frogs, toads, and tortoises are frequently used to demonstrate the heartbeat; observation is made easier by the low rhythm and relatively slow movements. Direct observation of the beating heart in homoiothermic animals is not so simple, as special methods are required, but it is of great interest because the results can be more directly applied to man. Rabbits, cats, and dogs are most frequently used. The animal is anesthetized, and the thorax is opened. Artificial respiration must be applied before the thorax is opened, by rhythmically inflating the lungs with a bellows or an air pump, and must be kept up throughout the experiment. The pericardium is opened, and by sewing it onto the chest, an ample breach is made which exposes the heart. Direct observation of the human heartbeat was first recorded by Harvey, in the now historic case of the Earl of Montgomery, who had a malformation that exposed the heart. In the course of thoracic operations, direct observations of the heart have also been made.

A semidirect observation can be made by means of x-rays. By radioscopy it is possible, without opening the thorax, to observe the cardiac shadow and its variations during the course of the heartbeat.

Visual or tactile inspection of the heart does not give much information, because of the short duration and the rapidity of the changes that take place. Graphic registration must be used for accurate and detailed observation. Many methods have been devised by the careful application of mechanical principles; these will be described in the course of the following chapters.

DIRECT OBSERVATION OF THE HEARTBEAT

When the mammalian heart is directly observed *in situ*, the most prominent features are the contraction of the auricular appendages during auricular systole and the contraction of the ventricles with a slight rotation toward the right along the longitudinal axis. The ventricular volume does not vary conspicuously when observed with the naked eye, nor are there any changes in color. There is a change in the consistency of the ventricular wall, which is more marked in the right ventricle because its walls are not so thick as those of the left; in diastole they are dilated and flaccid, but become rounded and hard in systole. The changes in consistency can be readily appreciated by grasping the organ. The movements and changes in consistency of the heart cause the apex beat, which will be described later.

To obtain a simple record of the contraction of the mammalian heart, two pairs of connected Marey tambours can be used; one pair is tied to the auricle and the other to the ventricle, so that when the heart contracts the pressure rises within the tambour. This rise of pressure is communicated by a rubber tube to the other tambour of the pair, which records the rise on a kymograph. This method gives information only as to the sequence of the auricular and ventricu-

lar systoles, their regularity, and if a time marker is added, the heart rate.

THE PHASES OF THE CARDIAC CYCLE

The method described above does not permit the registration in detail of all the events in the cardiac cycle. Accurate knowledge has been obtained by correlating several aspects of the activity of the heart (Fig. 23). The different records that have afforded a base to delimit and recognize the phases of the cardiac cycle will be considered in detail further on, but it is convenient to enumerate these phases here, as they will be frequently referred to in the following paragraphs.

The succession of coordinated movements that takes place in the heart from the beginning of an auricular systole to the beginning of the following auricular systole constitutes a cardiac cycle. The three most important components of the cycle are, in order of appearance, (*a*) auricular systole, or presystole; (*b*) ventricular systole; (*c*) ventricular diastole.

Each one of these divisions comprises several phases, which correspond to well-defined events in the activity of the heart. For special purposes these phases can be considered as made up of subphases.

Auricular systole or *presystole* has an average duration in man of 0.11 sec. It is frequently considered as the last phase of ventricular diastole. Auricular contraction is not essential for the maintenance of an adequate circulation. Many persons with auricular fibrillation live for years without necessarily suffering from circulatory insufficiency.

Ventricular systole begins immediately after auricular systole has ended. If a low heart rate prevails there may be a short pause, the so-called "intersystolic interval," between the contractions of the auricle and the ventricle. Two well-defined phases can be distinguished in ventricular systole. During the first phase ventricular pressure increases, but the contents of the ventricle are not altered because the semilunar valves are kept closed by the pressure in the arteries, and the auricular-ventricular (A-V) valves close almost immediately after the ventricle begins to contract. During the second phase intraventricular pressure reaches and rises above the pressure in the arteries; the semilunar valves are therefore opened and blood flows from the ventricles into the arteries. The first phase is called the *isometric contraction phase* (Greek ἴσος, equal, and μὲτρον, measure), because the ventricular volume remains constant. The second is called the *ejection phase*, because the ventricle evacuates its contents; it extends

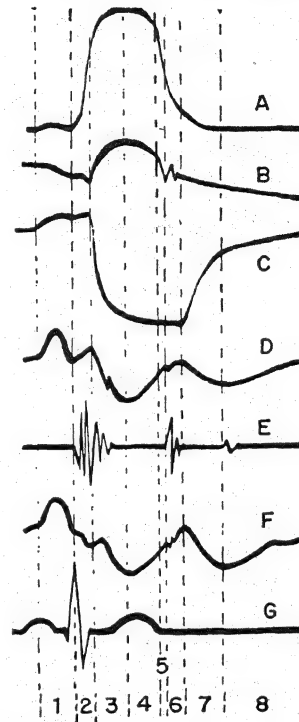



Fig. 23. Phases of cardiac activity. *A*, intraventricular pressure; *B*, aortic pressure; *C*, ventricular volume; *D*, intra-auricular pressure; *E*, phonocardiogram; *F*, phlebogram; *G*, electrocardiogram; 1, auricular systole or presystole; 2, isometric contraction; 3, maximum ejection; 4, reduced ejection; 5, protodiastole; 6, isometric relaxation; 7, rapid filling; 8, diastasis.

from the opening to the closure of the semilunar valves, which occurs as soon as the ventricles begin to relax.

Two subphases have been distinguished in the isometric contraction phase: the *entrant phase* and the *massive-contraction phase*. During the former, the fibers that constitute the ventricle enter successively into contraction and intraventricular pressure rises slowly; during the latter, when the excitatory state has spread to all the fibers, they are contracting simultaneously and intraventricular pressure rises rapidly. The whole isometric contraction lasts about 0.05 sec.

During the ejection phase, three subphases can be distinguished. Immediately after the semilunar valves

are opened there is a short period during which only a minimum amount of blood is ejected because of the inertia of the mass of blood in the large arteries (*minimal ejection subphase*). A subphase of *maximal ejection* follows, during which the greater part of the blood in the ventricles is discharged into the arteries.

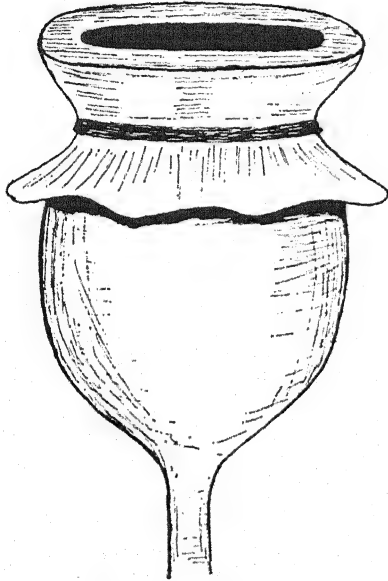



FIG. 24. Cardiometer.

Finally there is a subphase of *reduced ejection*. The whole ejection phase lasts 0.22 sec., the minimal eiection subphase 0.01 sec., maximal ejection subphase 0.08 sec., and reduced ejection subphase 0.13 sec.

Ventricular diastole commences when the ventricular musculature relaxes. It comprises four phases: protodiastole, isometric relaxation, rapid inflow, and retarded inflow or diastasis. The first effect of ventricular relaxation is a drop in pressure within the ventricle and the closure of the semilunar valves, marked by a sharp notch (incisura) in the aortic pressure curve. This *protodiastolic phase* lasts 0.04 sec. and its end coincides with the second heart sound. It is usual, therefore, to consider this sound as indicating the end of systole and the beginning of diastole.

Immediately after the closure of the semilunar valves, the A-V valves are still closed, and they remain so for about 0.08 sec. During this interval the ventricles relax without any significant change in volume. This phase is known as the *isometric relaxation*.

Following this phase the A-V valves open and blood accumulated in the auricles during ventricular systole, protodiastole, and isometric relaxation passes into the ventricles. At first a large amount of blood flows into the ventricle (*rapid inflow*), but later it flows at a lower rate, and continues to do so up to the end of the heart cycle. Rapid inflow lasts about 0.11 sec., and *retarded inflow*, or *diastasis*, about 0.19 sec.

In a resting subject, with a heart rate of 75 beats per minute, each cycle lasts about 0.8 sec.; ventricular systole has a duration of 0.3 sec., and ventricular diastole lasts 0.5 sec.

VENTRICULAR VOLUME

Detailed observation of the changes in ventricular volume gives considerable information on many aspects of cardiac function. It must be made with adequate recording apparatus, and preferably by optic registration.

Changes in ventricular volume are studied experimentally by means of the plethysmographic method, following François Frank's technique, in which the heart is introduced into a glass or metal vessel closed by a rubber diaphragm up to the A-V junction. The heart must fit snugly, but without being compressed by the rubber (Fig. 24). Changes in pressure caused in the plethysmograph by the heartbeat are transmitted through rubber tubing to a Marey recording tambour, or to a Frank segment capsule for optical recording. A rise in the curve records an increase in volume; a decrease in volume is recorded by a drop.

Even with the most perfect apparatus it is difficult to obtain ventricular volume curves completely free from artefacts. At the beginning of systole the whole heart has a tendency to penetrate farther into the cardiometer, and thus a rise in pressure is caused that does not correspond to a change in volume. Frequently the first part of the pulmonary artery is enclosed in the plethysmograph; thus, especially at the beginning of the ejection phase, an artefact caused by changes in the volume of this artery is introduced into the ventricular volume curve. Also during isometric relaxation, a small increase in ventricular volume is sometimes recorded, which is attributed to the entrance of blood into the intramuscular branches of the coronary arteries. As these artefacts occur at definite moments, by careful analysis their effects can be properly evaluated.

Ventricular volume is conveniently recorded simultaneously with aortic or auricular pressures so as to have points of reference in the analysis of the curves.

Cyclic changes in ventricular volume. In each cardiac cycle there are changes in volume due to the emptying and filling of the ventricles. Figure 25 reproduces ventricular volume and aortic pressure curves recorded simultaneously with a Frank segment capsule and a Wiggers optical manometer respectively. In the following description, the letters in parentheses refer to the segments marked on the curves.

FIG. 25. Simultaneous records of ventricular volume (upper tracing) and aortic pressure (lower tracing) in the dog. Time in 0.2 sec.: *a-b*, presystole; *b-c*, isometric contraction; *c-d*, maximum ejection; *d-e*, reduced ejection; *e-f*, isometric relaxation; *f-g*, rapid inflow; *g-a*, diastasis.

During auricular systole (*a-b*) the ventricular volume increases more or less according to the length of the preceding diastole; the longer the diastole, the smaller the amount of blood contributed by auricular systole to the filling of the ventricle. In a heart beating at a normal resting rate, this contribution is not very important. Neither is it necessary, as is proved by cases of auricular fibrillation with good circulatory conditions.

During isometric contraction (*b-c*), *i.e.*, between the beginning of ventricular systole and the opening of the semilunar valves, there are no changes in ventricular volume because all the heart valves are closed, so that blood can neither enter nor leave the ventricles. A small rise in the volume curve is nevertheless almost always recorded; this artefact is due to the tendency of the ventricle to penetrate farther into the cardiometer, as has already been mentioned.

During the ejection phase (*c-e*) ventricular volume decreases, at first rapidly, the tracing dropping almost vertically (*c-d*), and then more slowly, so that the tracing becomes almost horizontal (*d-e*). This decrease in volume is due to the ejection of blood into the arteries, at first rapidly (maximum ejection) and then at a lower rate (reduced ejection).

Between the closing of the semilunar valves and the opening of the A-V valves (*e-f*), *i.e.*, during isometric relaxation, when blood does not enter or leave the ventricle, the curve rises slowly. This is attributed to the entrance of blood into the coronary arteries within the ventricular walls, which are compressed during the systole by the contraction of the myocardium.

The opening of the A-V valves causes a rise in the curve, which at first takes place rapidly (*f-g*) and then more slowly (*g-a*). The ventricle fills in two phases: one of rapid inflow and then another of retarded inflow, or diastasis. Auricular systole follows the phase of diastasis and begins a new cycle.

The diastolic size of the heart is recorded at the end of diastole, *i.e.*, at the moment of maximum ventricular volume. The systolic size is recorded at the end of ventricular systole, *i.e.*, when the ejection phase has ended and the ventricular volume is at its minimum.

The mean ventricular volume. Besides the changes in ventricular volume that occur in the course of the cardiac cycle, it is interesting to note the variations in volume through a series of cycles, as in certain physiologic circumstances

the heart in the successive cycles does not have the same systolic or diastolic size. The mean ventricular volume must be considered when following these changes through several cycles. An idea of this mean volume can be obtained from the general direction of the plethysmogram. The mean ventricular volume is an abstract idea; it would correspond to the line joining the points representing the average volume of the ventricle during each cycle. Heart rate, venous return, and peripheral resistance to blood flow are factors that modify the mean cardiac volume. Changes in the heart rate cause variations in the duration of the diastole. As the heart rate increases, the filling time of the ventricles is shortened. If there is moderate tachycardia, the shortening of the diastole affects only the last moments of retarded inflow or diastasis; therefore the ventricular volume is only slightly reduced. If there is considerable tachycardia, shortening of diastole affects the phase of rapid inflow and thus impairs ventricular filling; the diastolic size decreases and consequently the mean ventricular volume diminishes. Bradycardia has the opposite effect on diastolic size and mean ventricular volume.

Changes in venous return produce parallel changes in the mean ventricular volume. An increase in venous return increases the mean ventricular volume, and vice versa.

Changes in the peripheral resistance, which must be overcome by the heart to eject the blood, also act on the mean ventricular volume. An increase in resistance increases ventricular volume; diastolic size is mainly responsible for the change. Variations in the mean ventricular volume are therefore approximately proportional to the variations in venous return and peripheral resistance; but it does not follow that there is a strictly linear relation between these variables.

Electrokymography. Fluorocardiography. If a photoelectric cell is applied on the margin of the shadow cast by the heart on the x-ray fluoroscope, it is possible to register the movements of the heart during the cardiac cycle. This procedure was first called electrokymography[1] and later fluorocard ography.[2] The current arising in the photoelectric cell is recorded by means of a string galvanometer such as is used in electrocardiography. The photoelectric cell is placed in a closed box, with a slit through which it is stimulated. It can be placed on any part of the margin of the shadow cast by the heart or the large arteries. Records obtained by placing the cell on the margins of the right and left auricles, the aorta, or the pulmonary artery are similar to the corresponding pressure records; those of ventricular movements are similar to volume records.[1] The procedure is not, however, exempt from artefacts.[2]

[1] HENNY, G. C., B. R BOONE, and W. E. CHAMBERLAIN, *Am. J. Roentgenol.*, **57**, 409, 1947.

[2] LUISADA, A. A., F. G. FLEISCHNER, and M. B. RAPPAPORT, *Am. Heart J.*, **35**, 336 and 348, 1948.

DIASTOLIC SIZE AND STRENGTH OF CONTRACTION

The law of the heart. One of the most remarkable features of cardiac physiology is the way in which the ventricle responds to an increase in its contents. This capacity to respond is of primary importance for the adaptation of the heart to different circulatory conditions. When the contents of the ventricle increase, whatever the cause, the cardiac fibers lengthen and there is a more vigorous contraction; intraventricular pressure increases more rapidly, isometric contraction is shortened, the ejection phase is lengthened, and the velocity of ejection increases. This response is known as Starling's law of the heart. As the load of the heart increases, so does the strength of its contraction, but only up to a critical limit, above which contractions become progressively weaker. In this condition of circulatory decompensation the heart is not sufficiently emptied at each systole, and it dilates progressively without a corresponding increase in the strength of its contraction. Catheterization of the human heart has shown that in man the right ventricle also responds to Starling's law of the heart.[3]

When ventricular dilatation occurs within the limits of the law of the heart, it is said to be a tonogenic dilatation (Moritz). When the capacity of adaptation of the heart has been surpassed and the myocardium

[1] LUISADA, A. A., and F. G. FLEISCHNER, *Proc. Soc. Exper. Biol. & Med.*, **66**, 436, 1947. MEDNICK, H., J B. SCHWEDEL, and P. SAMET, *Circulation*, **2**, 250, 1950; SALANS, A. H., J. A. SCHACK, and L. N. KATZ, *Circulation*, **2**, 900, 1950.

[2] ZINNGER, H. F., C. F. KAY, and J. M. BENJAMIN, *Circulation*, **2**, 197, 1950.

[3] LAUSON, H. D., A. COURNAND, and R. A. BLOOMENFIELD, *J. Clin. Investigation*, **25**, 913, 1946.

does not respond to an increase in load by a more vigorous contraction, it is said to enter into myogenic dilatation (Moritz).

Whether the greater length of the fiber, caused by the increased volume, is the cause of the more vigorous contraction, as Starling believed, or whether this is due to an increase in initial tension produced by the lengthening of the fiber, is a subject still under debate. This discussion is mainly of academic interest, because under ordinary conditions, whenever the fiber is lengthened its tension is increased.

Starling and his associates also found that an increase in the oxygen consumption of the myocardium accompanied the increased diastolic volume of the ventricles responsible for the more vigorous contraction. This metabolic increase, within certain limits, is proportional to the elongation of the fibers.

Hypertrophy of the heart. In certain conditions the heart can become totally or partially enlarged, because of a more or less considerable increase in the mass of its fibers.

Hypertrophy results from persistent diastolic overloading with moderate distention of the cardiac fibers, due to incomplete evacuation, as in cases of valvular stenosis or increased peripheral resistance, or to the reception of abnormal additional loads, as in cases of valvular insufficiency.

Moderate distention increases the strength of systole and is therefore the first compensatory factor in response to the abnormal condition. Probably it is also the initial stimulus of the metabolic process that produces cardiac hypertrophy. This hypertrophy is a more permanent compensatory mechanism, which restores to a certain extent the reserves that have been depleted.

Cardiac hypertrophy is prominent in certain pathologic conditions, but it is also a physiologic reaction. At birth both ventricles have approximately the same development, but when the pulmonary circulation is established and the ductus arteriosus and Botallo's foramen are closed, each ventricle has to fulfill a different task; the left ventricle must overcome a resistance far greater than that encountered by the right one. Gradually the left ventricle develops more than the right; in a normal adult its walls become four to five times as thick. The mechanism of this physiologic hypertrophy is probably the same as that of pathologic hypertrophy.

MYOCARDIOGRAPHY AND MYOCARDIOGRAMS

The contraction of the heart fibers can be studied in the heart beating *in situ*, or in the isolated heart, by means of a recording device known as a myocardiograph. Records so obtained are called myocardiograms. There are several types of myocardiograph, but each consists essentially in a pair of lever arms articulated together so that the shortening and lengthening of the cardiac fibers, to which they are attached, separates or approximates them; this movement then acts on a suitable recording system. The records obtained are frequently altered by artefacts, and their interpretation requires considerable experience.

Cinefluorographic angiocardiography. A substance which is opaque to x-rays, *e.g.*, a 70 per cent solution of diodrast, is injected into the jugular vein in dogs, and subsequent events are recorded by means of motion pictures of the fluoroscopic image of the heart, taken for periods of 10 to 14 sec. This procedure shows that diastolic filling of the ventricle takes place more rapidly than systolic emptying, and that a certain amount of residual blood remains in the ventricle at the end of systole. Auricular contraction produces little change in the diastolic size of the ventricle.[1]

A similar technique has been applied in man. This consists in injecting diodrast directly into the cavities of the heart by transthoracic puncture.[2]

THE APEX BEAT

If the anterior thoracic wall at the level of the fourth or fifth intercostal space, slightly medial to a vertical line passing through the nipple, is palpated, a sensation will be felt as if the thoracic wall were lifted at each heartbeat. This is known as the apex beat.

In pathologic conditions, owing to displacement, hypertrophy, or dilatation of the heart, the place and sometimes the nature of the apex beat are changed.

The sensation perceived is due to the sudden hardening of the myocardium and the changes in position caused by the ventricular systole (R. Tigerstedt).

[1] Rushmer, R. F., and N. Thal, *Circulation*, **4**, 219 1951.

[2] Ponsdomenech, E. R., and V. Beato-Nuñez, *Am. Heart J.*, **41**, 643, 1951.

The apex-beat cardiogram. The apex beat can be registered, and the record obtained is known as the apex-beat cardiogram. Marey was the first to obtain this type of record by means of a cardiograph (Fig. 26) connected by rubber tubing with a recording tambour. The form of the curve obtained depends on the location of the button of the cardiograph. If it is placed exactly over the pulsating area, the hardening of the apex is registered; the tracing rises above the base line, and is similar to an intraventricular pressure curve. If the button is not placed exactly above the pulsation, the curve obtained is the equivalent of a cardiac plethysmogram (volume curve), and falls mostly below the base line.

Fig. 26. Marey's cardiograph.

The apex beat can be recorded by the optical method. A small funnel, connected by rubber tubing with a Frank segment capsule, is placed over the apex and the variations in pressure are registered in a photokymograph. Tracings of optically recorded apex beats are given in Figs. 27, 28, and 29. The most important features are

1. An auricular wave due to the distention of the ventricle caused by the blood ejected into it by the auricular contraction.
2. The principal or systolic wave, which shows considerable variations in different individuals or even in the same person according to the location of the receptive funnel or the pressure exerted on it. In some instances it closely resembles an intraventricular pressure curve (Fig. 27); in others, the curve, after an ascending phase of variable length, shows a fall very similar to that of the ventricular volume curve during the ejection phase (Fig. 28). Finally, there may be a blending of both types of curves, giving rise to a considerable number of intermediate varieties (Fig. 29).
3. At the beginning of diastole, corresponding to the rapid inflow phase, there is a more or less clear and rapid rise in the curve, caused by the sudden distention of the ventricle by the blood that enters it when the A-V valves open.

If a very sensitive recording system is used, the vibrations of the heart sounds may appear superimposed on the curve described.

In an apex-beat curve it is possible to recognize auricular systole, the commencement of ventricular systole, sometimes the isometric contraction and ejection phases, the rapid inflow, and the retarded inflow or diastasis.

The apex-beat curve records mainly changes occurring in the left auricle and ventricle, while the venous pulse (to be described later) portrays mainly the activity of the right cavities. Simultaneous records of the apex beat and venous pulse will make it possible to determine the degree of coordination of the activity of the right and left ventricles, and so to establish which ventricle contracts first in cases of ventricular extrasystoles or bundle-branch block.[1]

CARDIOPNEUMATIC PHENOMENA

The thorax, with its semirigid walls, behaves as a plethysmographic chamber which registers changes in ventricular volume and the ejection of blood out of the thorax. There is a slight fall in intrathoracic pressure at each ejection phase of the ventricular systole, especially when the glottis, or the nose and mouth, are closed. These variations of intrathoracic pressure due to the activity of the heart are known as cardiopneumatic phenomena. They can be recorded by connecting the trachea by means of rubber tubing with a Marey tambour or a Frank segment capsule. The curve so obtained is a cardiac volume curve. The subject must give considerable cooperation to obtain good tracings.

FUNCTIONS OF THE PERICARDIUM

The pericardium is a serous membrane which separates the heart from the neighboring

[1] Battro, A., E. Braun Menendez, and O. Orías, *Rev. argent. de cardiol.*, 3, 325, 1937.

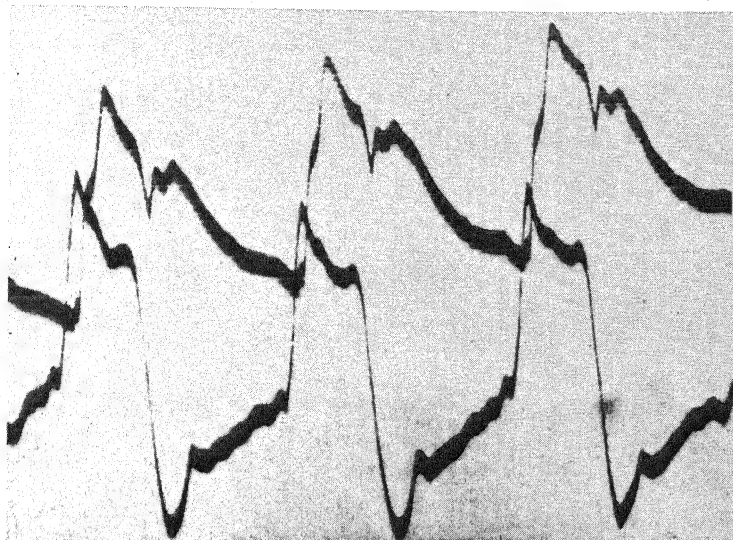



FIG. 27. Optical record of human carotid pulse (upper tracing) and apex-beat cardiogram (lower tracing). The apex-beat cardiogram in this case has registered the hardening of the apex and the record resembles that of intraventricular pressure. Note the delay of the carotid pulse with respect to the apex-beat cardiogram. (*Records obtained by Dr. E. Braun Menéndez.*)

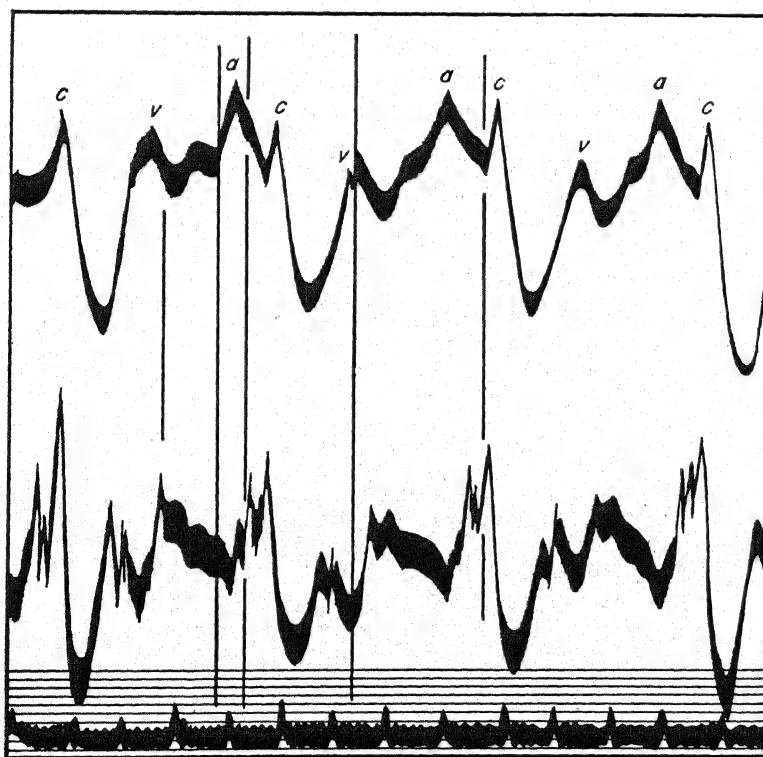



FIG. 28. Phlebogram and apex-beat cardiogram. The latter is the equivalent of a record of the heart volume. Note that the descending limb of the *v* wave in the venous pulse (rapid filling) coincides with a rise in the cardiogram. (*Records obtained by Dr. E. Braun Menéndez.*)

viscera. It forms a receptacle closely adapted to the cardiac surface, which is permanently lubricated, thus facilitating the movements of the heart.

The inelastic, resistant fibers of the pericardium protect the heart from sudden and excessive dilatation. The pericardium, on the other hand, can be considerably distended by a force of small intensity if it acts continuously. Thus the capacity of the pericardium can be considerably increased in cases of cardiac hypertrophy, progressive dilatation of the heart, and pericardial effusion.

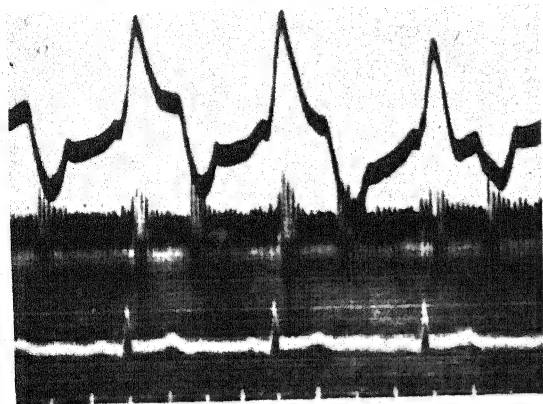



FIG. 29. Apex-beat cardiogram, phonocardiogram, and electrocardiogram. The apex-beat cardiogram in this case is of the intermediate type, recording first the hardening of the apex and then ventricular volume. (*Records obtained by Dr. E. Braun Menéndez.*)

The pericardium is not indispensable for the normal activity of the heart; it can be completely extirpated without causing any outstanding disturbance.

BIBLIOGRAPHY

FREY, W., in Bethe's "Handbuch der normalen und pathologischen Physiologie," **7**, 221, Springer, Berlin, 1926.

HENDERSON, Y., *Am. J. Physiol.*, **16**, 325, 1906 (ventricular volume).

HEYER, H. E., and B. R. BOONE, The Present Status of Electrokymography, *Am. Heart J.*, **44**, 458, 1952.

PACHON, E., and R. FABRE, "Le Cardiogramme de décubitus latéral gauche en clinique," Gaston Doin, Paris, 1929.

REGELSBERGER, H., *Ztschr. f. klin. Med.*, **104**, 107, 1926 (apex beat).

STEINHAUS, A. H., *Physiol., Rev.*, **13**, 103, 1933 (cardiac hypertrophy).

TIGERSTEDT, R., "Physiologie des Kreislaufes," Walter de Gruyter & Company, Berlin and Leipzig, 1921 (cardiograms, movements and volume of the heart, etc.).

WIGGERS, C. J., "Modern Aspects of Circulation in Health and Disease," Lea & Febiger, Philadelphia, 1923.

WIGGERS, C. J., "The Pressure Pulses in the Cardiovascular System," Longmans, New York, 1928.

WIGGERS, C. J., "Physiology in Health and Disease," 4th ed., Lea & Febiger, Philadelphia, 1944.

CHAPTER 13

Pressure Changes in the Heart

CHAUVEAU AND MAREY[1] were the first to obtain records of pressure changes in the heart. Their observations were made on the horse by means of special catheters.

The right auricle and ventricle were explored with a single flexible catheter, the walls of which could not be flattened; at one end it had two rubber balloons sustained by a metal frame. Each one of these balloons was connected separately by rubber tubing with a Marey tambour, which registered the pressure changes on the smoked surface of a kymograph. The catheter was passed down the jugular vein so that the balloon on its end was placed in the right ventricle, and the more proximal balloon in the right auricle.

The left ventricle was explored by means of a metal catheter ending in a rubber balloon, which was passed through the carotid artery and the aorta into the ventricle. This catheter was connected with another Marey tambour. Figure 30 shows a diagram of the apparatus, and Fig. 31 the curves obtained.

Chauveau's and Marey's work is of great historical importance, as it inaugurated an era in the physiology of the circulation. Their results were soon surpassed by those of other workers who used more accurate methods. Their apparatus did not give a faithful record of all the pressure changes, because it could not respond with sufficient speed owing to its inertia; moreover, friction on the recording surface damped the pressure waves and suppressed significant details in the curve.

[1] CHAUVEAU, A., and E. J. MAREY, *Gaz. méd. Paris*, p. 675, 1861. *Compt. rend. Acad. d. sc.*, **53**, 622, 1861; **59**, 32, 1862.

Frank's theoretical and experimental studies led to the construction of instruments of great efficiency. The replacement of levers by a beam of light permitted the use of capsules with very tensely drawn membranes without diminishing the sensitiveness of the apparatus; these responded rapidly, were practically aperiodic, and had no inertia. Frank's optical manometer was later improved by Wiggers, who has applied the method widely in the study of the pressure pulses in different parts of the cardiovascular system.

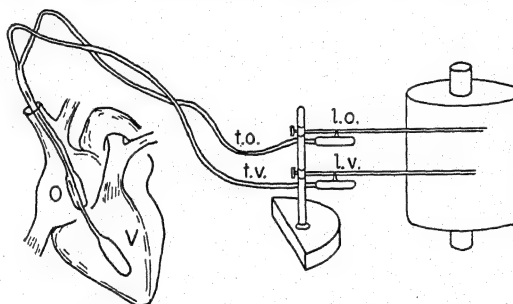



FIG. 30. Diagram of Chauveau's and Marey's apparatus for registering intracardiac pressure changes.

PRINCIPLES OF MANOMETRIC REGISTRATION

According to Wiggers[1] the requirements of a good manometer are (*a*) that it inscribe a curve that has a sufficient amplitude to show changes in gradient and finer vibrations; (*b*) that it respond without measurable lag; (*c*) that it reproduce the pressure variations correctly as regards amplitude and phasic relations.

[1] WIGGERS, C. J., "The Pressure Pulses in the Cardiovascular System," Longmans, New York, 1928.

The amplitude of the curve depends on the sensitivity of the instrument; *i.e.*, on the relation between the magnitude of its response and the force that provokes it. The sensitivity of an instrument is greater as the force necessary to provoke a response of a certain magnitude is smaller.

The efficiency of an instrument in recording the correct amplitude and phasic relations of pressure changes depends on its dynamic properties—inertia and natural frequency—which condition its capacity to begin and to cease registering without delay.

Recording manometers are periodic systems, and if their dynamic attributes are not adapted to the pressure changes which they must record, the resulting curves will be vitiated by faults due to the inefficiency of the instrument.

To obtain a faithful record the natural frequency of the recording system should be at least five times the frequency of the shortest component of the variation to be registered.

Otto Frank analyzed and correlated mathematically the factors that condition sensitivity and natural frequency of registering systems, and he was able to develop a firm theoretical basis for the construction of manometers of great fidelity. Thus the segment capsule came into use. In this the registering lever is replaced by a beam of light which has no inertia and is absolutely rigid. Frank's segment capsule plays in optical registration the part played by Marey's tambour in the registration on a smoked surface. It is used for optical registration not only in connection with manometers but also, as will be seen later, in other recording systems.

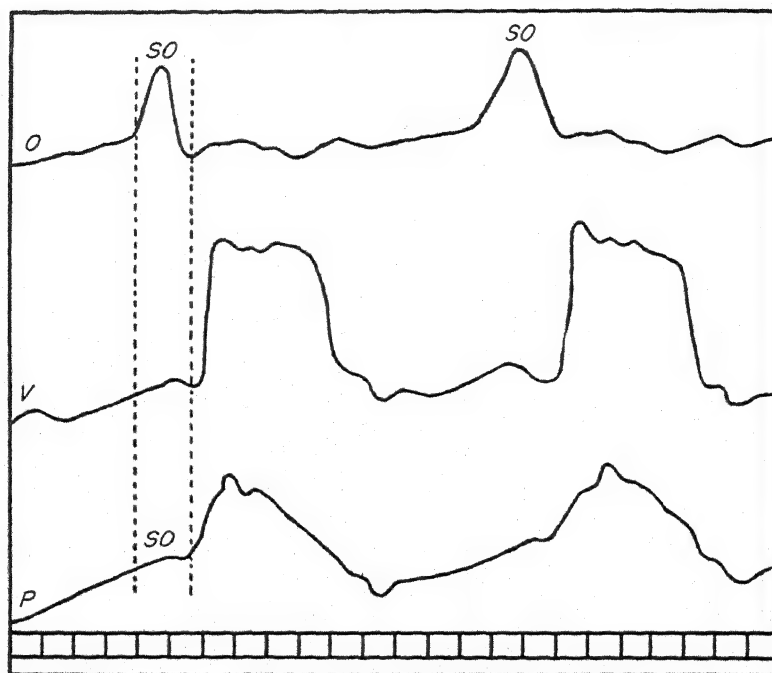



FIG. 31. Cardiac pulses in the horse. O, right auricle; V, right ventricle; P, apex-beat cardiogram; SO, auricular systole (*Chauveau and Marey.*)

Frank's segment capsule is a small hollow cylinder, one end of which is covered by a rubber membrane at a suitable tension, on which a small mirror is attached. The diameter of the cylinder is 7 to 9 mm., and its height less than 1 cm. (Fig. 32). The end covered by the rubber membrane is flattened on one side, so if seen "end on" it appears like a circle from which a segment has been cut off; hence the name of segment capsule. The mirror is attached to the membrane so that one of its edges coincides with the flattened side of the capsule, which thus acts as an axis for the movements of the mirror when the membrane swells out or is depressed by the pressure changes within the system. As the mirror moves, the direction of a beam of light reflected on it changes, and these changes can be recorded on a photosensitive surface that moves in a direction at right angles to that in which the light beam moves. The apparatus that holds and moves the photographic film is called a photokymograph.

The natural frequency of a registering manometer obeys the following law, established by Frank:[1]

$$N = \frac{1}{2\pi}\sqrt{\frac{E'}{M'}}$$

The natural frequency (N) varies inversely to the square root of the effective mass of the system (M'),

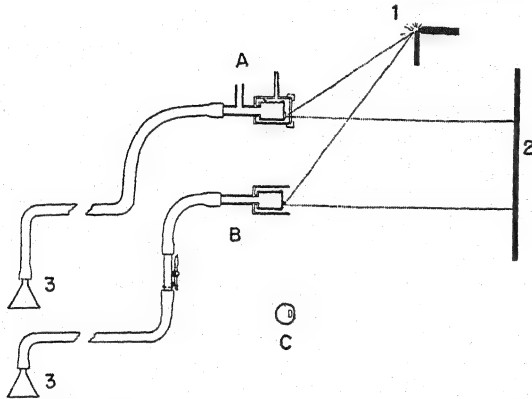



FIG. 32. Diagram of capsules of Frank and of Wiggers and Dean for optical registration. *A*, longitudinal section of Wiggers' and Dean's capsule for registering the heart sounds; *B*, longitudinal section of Frank's capsule for optical registration; *C*, end view of Frank's segment capsule and the situation of the mirror on the membrane; 1, arc lamp; 2, photographic film; 3, receiving funnel.

and directly with the square root of the volume-elasticity coefficient of the system (E'). Therefore, the natural frequency is increased by diminishing M' (large transverse section and short length of the system, high density of the transmitting fluid) or by increasing E' (diminishing distensibility: tense membranes, nondistensible material for the manometer and connections).

Frank's and Wiggers' manometers. Frank, and later Wiggers, following the principles established by Frank, endeavored to increase the natural frequency of manometers by reducing the effective mass (M'). The instruments constructed have great efficiency and give a faithful record of pressure changes in the circulatory system. Wiggers' universal optical manometer (Fig. 33) is very practical, but to insert the wide and short cannulas into the cavities of the heart or the arteries a rather traumatic operation must be performed.

Hamilton's manometer.[2] Hamilton and others have tried to increase the natural frequency by increasing the value of E' in Frank's law. A tensely drawn membrane of a very slightly distensible material allows the increase of the effective mass without diminishing the natural frequency. In this way smaller cannulas, even hypodermic needles, can be used together with a more or less long connecting system, which must be of nondistensible material, such as lead tubes. The decrease in sensitivity caused by increasing the tension of the membrane (beryllium membranes are used in current models) is compensated by lengthening the beam of light (increasing the distance between the manometer and photokymograph), a circumstance that requires a stronger source of light. A diagram of Hamilton's manometer is also given in Fig. 33. This instrument permits the recording of pressure variations in human arteries. It may also be used in connection with a catheter introduced into the heart chambers.

Gauer's manometer. When pressure pulses are registered with an intracardiac catheter, the effective mass is increased by the fluid column that fills the catheter. This diminishes the efficiency of the manometer, and the models mentioned above do not give satistactory results. Gauer and Gienapp[1] have designed a very small manometer, 2 cm. long by 3 mm. in diameter, which is easily adapted to the internal (intracardiac) end of the catheter. A diagram of this manometer is given in Fig. 34. Intracardiac pressure exerts its effect on a piston *A* against a spring *C*. Displacement of the piston makes its end (of soft iron) act on the coils of a differential transformer contained in the small compartments *D*. Cables from this transformer (not represented in the diagram) connect it through an amplifier with a recording galvanometer. A thin rubber diaphragm *B*, maintained in its place by a ring, prevents blood from entering into the system. The natural frequency of the manometer is 1,000 c.p.s. and it has a linear response from −50 to +250 mm. Hg. It can be easily calibrated with a mercury manometer by applying suction to the external end of the catheter, without having to remove it from the heart. Excellent records can be obtained with this instrument.[2]

Piezoelectric instruments. Recently instruments have come into use in which variations in pressure are converted into variations in an electric circuit (piezoelectric systems; πίεσις, pressure), which can

[1] FRANK, O., *Ztschr. f. Biol.*, **89**, 274, 1929.

[2] HAMILTON, W. F., G. BREWER, and I. BROTMAN, *Am. J. Physiol.*, **107**, 427 and 436, 1934; **112**, 130, 1935; *J. A. M. A.*, **107**, 853, 1936; GREGG, D. E., R. W. ECKSTEIN, and M. H. FINEBERG, *Am. J. Physiol.*, **118**, 399, 1937.

[1] GAUER, O., and E. GIENAPP, *Science*, **112**, 104, 1950.

[2] ELLIS, E. J., O. H. GAUER, and E. H. WOOD, *Proc. Staff Meet., Mayo Clin.*, **25**, 49, 1949; *Am. J. Physiol.*, **159**, 568, 1949; *Circulation*, **3**, 390, 1951.

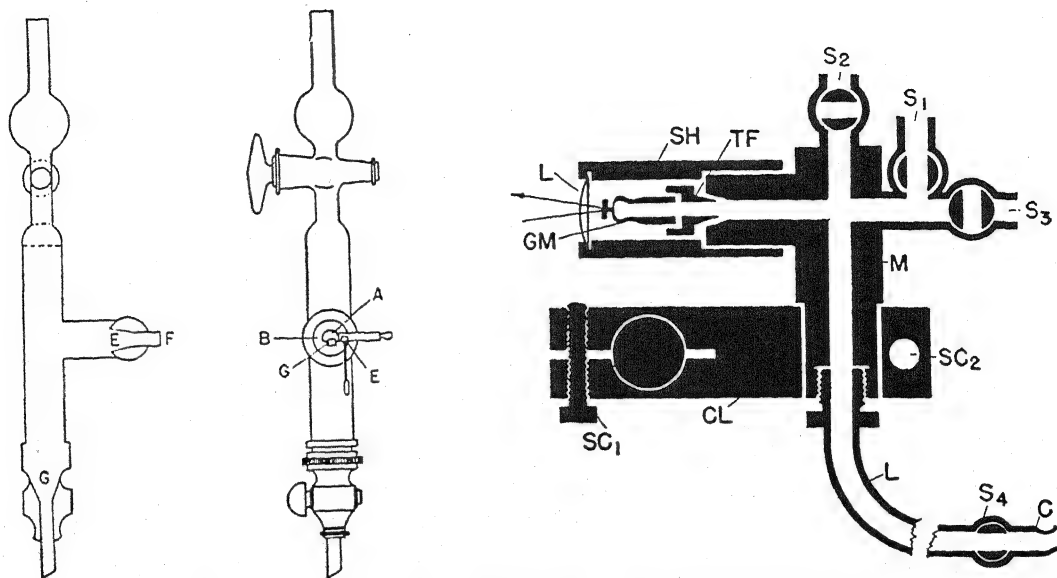



Fig. 33. Left: Diagram of Wiggers' manometer (front and lateral views). Right: Green's modification of Hamilton's manometer. (*Wiggers, C. J., "Pressure Pulses in the Cardiovascular System," Longmans, New York, and Green, H. D., Circulation. Physical Principles, in "Medical Physics," ed. by O. Glasser, vol. 1, p. 208.*)

be registered by appropriate galvanometers. Some of them consist in a Wheatstone bridge,[1] which varies in resistance.

As yet there is not sufficient experience to judge the efficiency and usefulness of these systems. Instruments most commonly in use are manufactured by the Statham Laboratories, Los Angeles, Calif., and by the Sanborn Company, Boston, Mass.

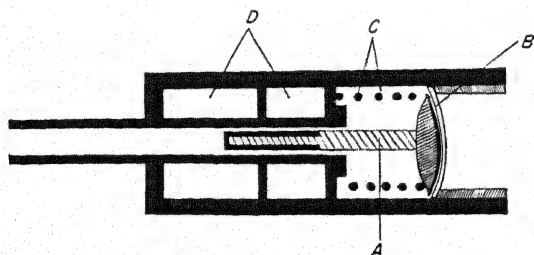



Fig. 34. Diagram of Gauer's manometer. (Description in text.)

INTRAVENTRICULAR PRESSURE CHANGES

The following analysis of the pressure pulses in the cavities of the heart is based on curves free of artefacts obtained by means of high-efficiency manometers. Intraventricular pressure should be recorded simultaneously with some other manifestation of cardiac activity, so as to have points of reference for the interpretation of the facts observed. Figure 35 reproduces intraventricular and aortic pressure records simultaneously obtained in a dog with Wiggers' optical manometers. The upper tracing corresponds to the aorta and the lower one to the left ventricle; the vertical lines in the lowest part of the record are separated by intervals of 0.04 sec.

Ventricular systole. Ventricular contraction begins at *L*. The pressure rises relatively slowly and at increasing speed up to *M*, whence it rises in a thin, almost straight line up to *N*. This point corresponds to the opening of the semilunar valves, as is shown by the fact that aortic pressure begins to rise precisely at this moment. From this moment onward, ventricle and aorta form one common cavity and the pressure changes portrayed by both curves are identical. The curves are more or less rounded, peaked, or flattened, according to the way pressure varies and according to the speed of the photokymograph. At *R* the semilunar valves close, and the ventricular pressure falls rapidly down to *S*, from which point onward it falls more gradually until it reaches the initial level.

[1] Lambert, E. H., and E. H. Wood, *Proc. Soc. Exper. Biol. & Med.*, **64**, 186, 1947; Motley, H. L., A. Cournand, L. Werko, D. Dresdale, A. Himmelstein, and W. R. Dickinson, Jr., *Proc. Soc. Exper. Biol. & Med.*, **64**, 241, 1947.

The ventricular pressure curve seldom falls below the initial level (negative pressure). A small rise *KL* is sometimes observed just before the ventricular systolic rise; it is due to the slight increase in pressure caused by the auricular systole.

Isometric contraction. The first part of the curve (*LN*) corresponds to the isometric contraction; the energy set free is entirely used to increase the tension of the heart fibers and the pressure in the ventricle, without causing any shortening of the fibers (some energy, of course, is lost as heat). Ventricular volume records show no change in size during this period; hence the term "isometric" (see page 97). As the A-V and semilunar valves are closed, blood does not enter or leave the ventricle.

At first the pressure increases gradually (*LM*) and then at greater speed (*MN*); it is usually accepted that this is due to the fact that, as excitation spreads through the myocardium, the heart fibers enter successively into contraction (*entrant contraction subphase*), and later when all the fibers are excited they are all contracted (*massive contraction subphase*).

Ejection phase. When ventricular pressure reaches the level of the aortic pressure and rises above it, the semilunar valves are opened and the mechanical conditions of ventricular contraction are altered. Energy of contraction is now used to eject the blood from the ventricle into the arteries (*ejection phase*). The heart performs "work," in the mechanical sense of the term. Immediately after the semilunar valves open (*NO*) most of the contraction energy is used to overcome the resistance of the column of blood in the aorta, and very little in moving this column (*minimal ejection subphase*). Between *O* and *P* ejection is at a maximum, and then again diminishes down to the end of systole *Q* (*reduced ejection subphase*).

Ventricular diastole. The segment *QR* corresponds to what is called the aortic incisura, or protodiastolic phase; at *R* the semilunar valves close.

From *R* onward the semilunar and A-V valves remain closed. The contents of the ventricle cannot vary, so ventricular relaxation is not accompanied by any change in the length of the fibers (*isometric relaxation*). The A-V valves open at *S* and blood flows into the ventricles. The subphases of ventricular diastole are not portrayed in the ventricular pressure curve; they can be seen in the auricular pressure, jugular pulse, and ventricular volume curves.

Right ventricle. Pressure changes in the right ventricle are essentially the same as those in the left ventricle; details similar to those described above are found in the right ventricular

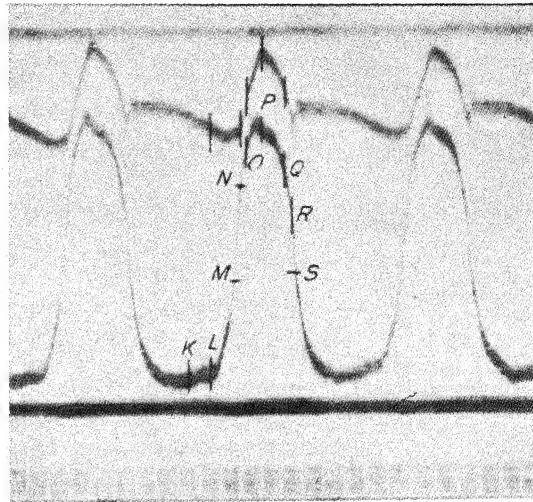



FIG. 35. Intraventricular pressure (lower record) and aortic pressure (upper record) curves. Time in 0.04 sec. *K-L*, auricular wave or presystole; *L-N*, isometric contraction; *L-M*, entrant contraction; *M-N*, massive contraction; *N-O*, minimum ejection; *O-P*, maximum ejection; *P-Q*, reduced ejection; *Q-R*, protodiastole; *R-S*, isometric relaxation. From the opening (*N*) to the closure (*R*) of the semilunar valves the intraventricular and aortic pressure changes correspond perfectly.

pressure curve. The only difference is that the right ventricle has a lower maximum pressure, corresponding to the thinner wall and less developed musculature of this ventricle. Simultaneous records of right and left ventricular pressures (Fig. 36), besides the difference in amplitude, show that the rise in pressure in the right ventricle precedes by a very short time the rise in the left ventricle. This very slight asynchronism is not constant and is never greater than a few milliseconds.

Chauveau's and Marey's curves, obtained in the horse (Fig. 31), differed in several respects from the curves obtained with optical manometers in dogs. Curves obtained with optical manometers in horses,[1] however, do not differ essentially from those of the dog obtained with the same method.

[1] HOUSSAY, B. A., O. ORÍAS, and L. GIUSTI, *Rev. Soc. argent. de biol.*, **12**, 259, 1936; *J. de physiol. et de path. gén.*, **34**, 1125, 1936.

Intraventricular pressure in man. Intraventricular pressure can be recorded in human beings by introducing a catheter through a vein in the arm into the right ventricle and connecting the catheter with a suitable manometer, *e.g.*, Hamilton's.[1] The curves do not differ significantly from those obtained in animals, but as there is a long column of fluid in the catheter, artefacts are frequently recorded. Right and left intraventricular curves have been obtained in man by direct transthoracic puncture.[2] The curves are similar to those of dogs, but the procedure is, of course, an exceptional one.

Left ventricle

Right ventricle

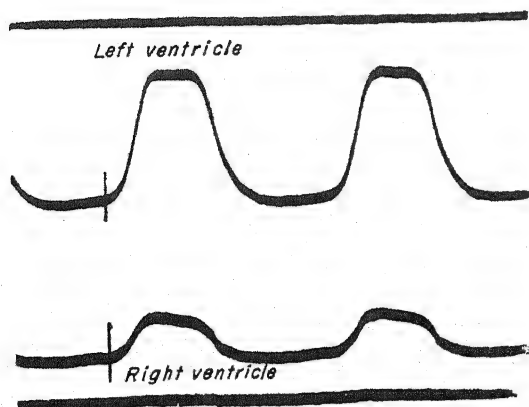



Fig. 36. Simultaneous records of left (upper tracing) and right (lower tracing) intraventricular pressure curves. Both manometers had the same sensitivity; the difference in amplitude is due to real differences in the pressures in the right and left ventricles.

INTRA-AURICULAR PRESSURE CHANGES

The pressure in the auricles is much lower than in the ventricles or the aorta; therefore the manometers used should be more sensitive than those used for registering ventricular or aortic pressures. With the sensitiveness of the membrane thus increased, it becomes almost impossible to avoid the recording of other phenomena of cardiac activity, such as the heart sounds that are superimposed on the pressure curve.

The pressure curve of the left auricle of a dog (essentially the same as the right auricular pressure curve) is reproduced in Fig. 37. It has been obtained simultaneously with a left intraventricular pressure curve by means of Wiggers' optical manometers.

Auricular systole (presystole) produces a more or less rounded wave (Fig. 37, 1). Coincident with the isometric contraction of the ventricle there is a sharp rise in auricular pressure (Fig. 37, 2), probably due to the closure of the A-V valves, and a series of vibrations which correspond to the first heart sound. When the ejection phase begins, intra-auricular pressure drops suddenly, because the shortening of the ventricular fibers, as blood is ejected into the arteries, produces mainly a lowering of the base of the ventricles, while the apex remains almost stationary. This fall in intra-auricular pressure produces a slight suction of blood from the veins. Therefore while the ventricle is pumping out blood into the arteries, the auricle is sucking it in from the veins. The ejection impulse, however, is much greater than the sucking effect; the latter has little or no functional significance, while the pumping out of blood from the ventricle is the most important fact of cardiac activity in circulatory dynamics.

1 2 3 4

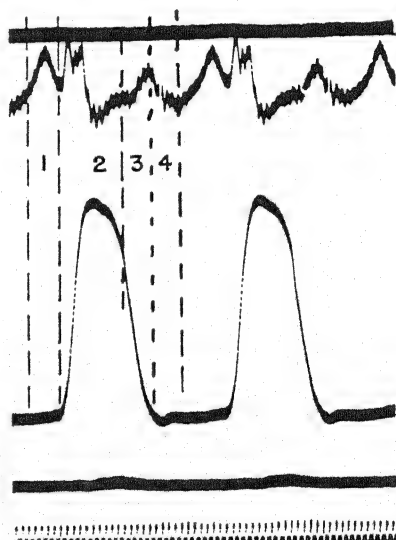



Fig. 37. Left intra-auricular (above) and intraventricular pressures registered by Wiggers' optical manometers. Time in 0.02 sec. 1, auricular systole; 2, isometric contraction and ejection phases; 3, isometric relaxation; 4, rapid inflow. The rise at the commencement of isometric contraction in the auricular pressure curve corresponds to the closure of the A-V valves.

During the reduced ejection phase there is a gradual increase in intra-auricular pressure,

[1] Cournand, A., *Bull. New York Acad. Med.*, **23**, 27, 1947.

[2] Buchbinder, W. C., and L. N. Katz, *Proc. Soc. Exper. Biol. & Med.*, **71**, 673, 1949.

which continues and becomes even more marked during the isometric relaxation of the ventricle (Fig. 37, 3). This rise in pressure is due to the continuous arrival of blood from the veins, blood that cannot pass into the ventricle because the A-V valves are closed. When these valves open, at the end of the isometric relaxation, blood flows into the ventricles and the pressure in the auricle falls (Fig. 37, 3-4). A wave is thus produced, the ascending limb of which is due to the accumulation of blood in the auricle (stasis) and the descending limb to its evacuation into the ventricle (rapid inflow phase). The summit of this wave corresponds to the opening of the A-V valves. After the ventricle has been almost completely filled, blood coming from the veins accumulates in the auricles, and the intra-auricular pressure rises (retarded inflow) up to the end of the cycle.

These changes in intra-auricular pressure are the main cause of the venous pulse, which will be studied later.

THE FUNCTIONING OF THE VALVES

The heart valves are membranes, made up of elastic connective tissue, which serve to direct the blood stream flowing through the heart. The movements of these valves are completely passive; they are caused by differences in the pressures of the cavities which they separate. They close the orifices on which they are inserted, and in normal conditions their efficiency is perfect. The auriculoventricular (A-V) valves (mitral valves on the left, and tricuspid valves on the right side of the heart) are opened when the pressure is greater in the auricles than in the ventricles and are closed when intraventricular pressure rises above that in the auricle. The semilunar valves in the aorta and pulmonary artery open when intra-ventricular pressure is higher than the corresponding arterial pressure, and close when the arterial pressure is the higher.

During the greater part of ventricular diastole, *i.e.*, from the end of isometric relaxation, throughout the auricular systole up to the beginning of ventricular systole, the A-V valves are open and blood flows into the ventricles. During this interval the semilunar valves remain closed. When ventricular systole commences, the A-V valves are suddenly closed, and a brief period elapses before the semilunar valves are opened (isometric contraction). During the ejection phase the semilunar valves are open while the A-V valves remain closed. The semilunar valves are closed in the protodiastolic phase, before the A-V valves open, so that during the isometric relaxation all the valves are closed.

The papillary muscles and the chordae tendineae, which arise from these muscles and are inserted on the free edges of the A-V valves, prevent the inversion of the valves when they close. The semilunar valves are disposed in a different way and close a smaller orifice, so they are not inverted in spite of not having any retaining fibers.

Inflammatory processes and congenital malformations can cause alterations in the valves, which disturb their function. The valves can thus narrow the orifices (valvular stenosis) or become incompetent (valvular insufficiency). In the first case, they hinder the passage of blood; in the second, part of the blood flows backward into the chamber from which it has been ejected. If these disturbances are very severe, they may be incompatible with life.

BIBLIOGRAPHY

FRANK, O., in Tigerstedt's "Handbuch der physiologischen Methodik," **2**, 1, S. Hirzel, Leipzig, 1913 (theory of optical manometers and the segment capsule).

MORITZ, F., in Bethe's "Handbuch der normalen und pathologischen Physiologie," **7**, 158, Springer, Berlin, 1926 (cardiac valves).

ORÍAS, O.: "Registro e Interpretación de la Actividad Cardiaca," 3d ed., El Ateneo, Buenos Aires, 1939.

STRAUB, H., in Abderhalden's "Handbuch für biol ogischen Arbeitsmethoden," **4**, 1, Urban & Schwarzenberg, Berlin, 1923 (theory of registration manometers).

WIGGERS, C. J., Determinants of Cardiac Performance, *Circulation*, **4**, 485, 1951.

WIGGERS, C. J.: "Modern Aspects of Circulation in Health and Disease," Lea & Febiger, Philadelphia, 1923.

WIGGERS, C. J., "The Pressure Pulses in the Cardiovascular System," Longmans, New York, 1928.

CHAPTER 14

The Heart Sounds

The activity of the heart produces certain sounds of low intensity, which can be heard by placing the ear directly on the chest wall, in the neighborhood of the heart, or by using a stethoscope for the same purpose.

Harvey, in his famous book on the movements of the heart and blood, mentioned the heart sounds, but Laënnec was the first to study these sounds and to point out the value they and their modifications had in medical practice. In 1819 Laënnec published a classic treatise on auscultation.[1]

Fig. 38. Stethoscopes.

The heart sounds are due to the vibrations that the activity of the heart causes in the heart itself or in neighboring structures. An untrained ear is not in a condition to explore the heart sounds profitably. The ear must be patiently trained, by methodic and prolonged practice; this training is a fundamental part of medical education.

Stethoscopes and phonendoscopes (Fig. 38) are used to aid the ear, especially in the localization of the sounds, and to reinforce them.

THE NORMAL HEART SOUNDS

In the course of the heart cycle at least two sounds are heard in all subjects. These stand out clearly above all others, and until recently they were the only ones generally recognized. Because of the order of their occurrence in the heart cycle, they have been called the first and second sounds. There is a short silence between them, and between the second sound of one cycle and the first sound of the following cycle there is a longer silent interval.

The first sound commences at the same time as the ventricular systole; it is described as being the lower in pitch of the two, the softer in quality, and the more prolonged. It lasts throughout the isometric phase and part of the ejection phase of ventricular systole. The second sound, which occurs practically at the end of the systole, is said to be sharper, shorter, and of a higher pitch than the first.

In a certain number of perfectly healthy subjects a well-trained ear can hear, immediately after the second, the so-called "physiologic third sound." Attention was clearly drawn to it for the first time by Gibson[1] in 1907.

Furthermore, in some normal subjects a very weak sound may be heard immediately before the first sound; it therefore precedes the ventricular systole. This sound is of such low intensity that it will not be perceived unless special care is taken to recognize it. The third sound and this other sound, the significance of which will

[1] Laënnec, T. H., "De l'auscultation médiate," Bosson et Chaudé, Paris, 1819.

[1] Gibson, A. G., *Lancet*, **2**, 1 and 380, 1907.

be examined later, frequently pass unnoticed on direct auscultation, especially if the observer ignores the possibility of their existence, but they are clearly marked in graphic records made with instruments of adequate sensitivity (Fig. 39).

The cause of the heart sounds. There are therefore four sounds per beat when the heart is contracting normally. Two of them are easily recognized; they are louder and more important than the other two, which are weaker and may not be noticed unless special care is taken during auscultation.

A detailed discussion of all the opinions as to the cause of the heart sounds would take up much space and would not be very profitable. Therefore only those which have a more solid foundation will be briefly considered. More details can be obtained in the monograph by Orías and Braun Menéndez.[1]

The first heart sound is due to four principal factors: (*a*) muscular contraction and tension of the walls of the ventricles (muscular factor); (*b*) closure of the A-V valves (valvular factor); (*c*) distention of the aorta and pulmonary artery at the beginning of the ejection phase (vascular factor); (*d*) residual vibrations due to auricular systole (auricular factor). The most important are undoubtedly the valvular and vascular factors. The others merely contribute to modify the fundamental character of the sound.

The second heart sound is due to the closure of the aortic and pulmonary semilunar valves. Therefore it takes place at the beginning and during the initial moments of isometric relaxation.

The third physiologic heart sound coincides with the end of the rapid filling phase, as has been demonstrated by the use of graphic registration. The moment of the cardiac cycle at which it occurs supports the theory that attributes it to vibrations of the ventricular wall, caused by the sudden distention produced by the rapid entrance of blood accumulated in the auricles and veins during the time the A-V valves are closed. Other theories as to the cause of the third sound are not supported by valid arguments.

The physiologic auricular sound is probably due to vibrations of (*a*) the auricular walls; (*b*) the A-V valves; (*c*) the ventricular walls distended by the auricular systole. Graphic registration has shown that vibrations originated in the course of auricular systole last longer than this systole, and the final oscillations are added to those of the first sound.

[1] ORÍAS, O., and E. BRAUN MENÉNDEZ, "The Heart Sounds," Oxford, New York, 1939.

Auscultation areas. The first heart sound is best heard over the lower part of the anterior chest wall. The clinical exploration of this sound gives valuable information on the functioning of the A-V valves, as closure of these valves is one of the main factors which produce it. The mitral valve is explored by auscultation over the apex of the heart (mitral area) and the tricusp over the base of the xifoid of the sternum (tricuspid area).

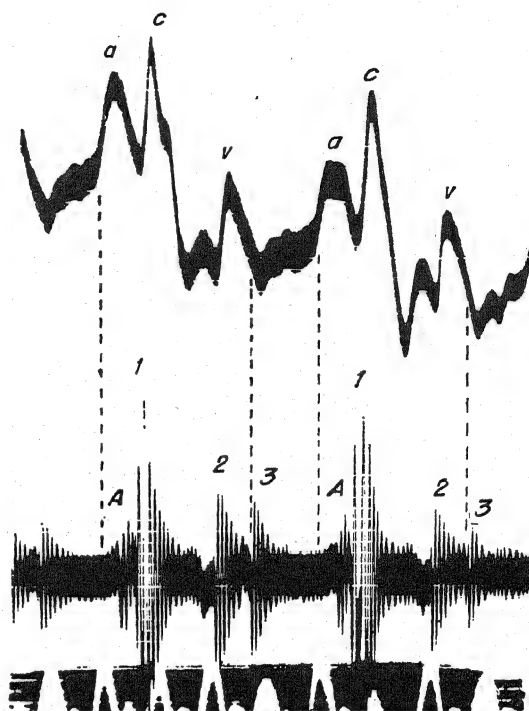



FIG. 39. The four normal heart sounds. Phlebogram, phonocardiogram, and time in 0.2 sec. The auricular sound *A* begins about 0.04 sec. after the commencement of the *a* wave of the venous pulse. The third sound coincides with the lower third of the descending limb of the *v* wave (end of rapid filling).

The second sound is best heard over the upper part of the anterior chest wall, and it gives information on the functioning of the semilunar valves. The aortic valves are explored by auscultation over the second right intercostal space, close to the sternum (aortic area), and the pulmonary valves over the second left intercostal space near the sternum (pulmonary area).

Graphic registration of the heart sounds. Phonocardiography. The vibrations which constitute the heart sounds can be registered graphically. The record obtained is called a phonocardiogram; it allows detailed analysis of the sounds, the moment in the heart cycle in which they occur, their duration, their number, and the amplitude and frequency of the vibrations.[1] The phonocardiogram is a valuable and permanent record.

Graphic registration of the heart sounds is based on the following general principle: the vibrations that would act on the tympanic membrane of the ear are made to act on an adequate resonator, the movements of which can be registered either directly, by transmitting them to a registering lever or to a beam of light, or indirectly, by making them modify an electric circuit, the oscillations of which are registered by a galvanometer. There are therefore two types of methods: (*a*) the direct or purely mechanical; (*b*) the indirect or electric.

Improvements in microphones, amplifiers, and recording galvanometers have greatly contributed to the advancement of electric phonocardiography. The records can be "picked up" by a photoelectric cell and again transformed into sound by ear receivers or loud-speakers.

[1] Phonocardiographs can omit, deliberately or otherwise, the registration of vibrations of certain frequencies.

Whatever the method used, it is convenient to register at the same time as the phonocardiogram, and on the same film, some other aspect of the heart's activity to have points of reference for the correct localization and recognition of the sounds. The venous pulse, the central arterial pulse, and the apex beat are especially useful for such a purpose. If an electrocardiogram is also simultaneously recorded, a complete picture of the heart's activity is obtained, but the electrocardiogram alone is inadequate as a reference record because it is silent during the greater part of ventricular diastole.

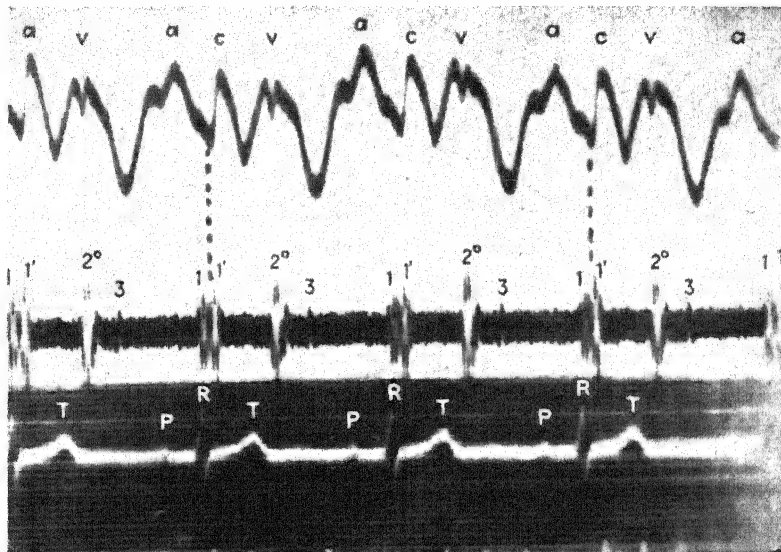



FIG. 40. Components of the first sound. Phlebogram, phonocardiogram, electrocardiogram, and time in 0.2 sec. The first sound is made up of two sets of vibrations, 1 and 1'. The vertical dotted line passing between the two components crosses the venous pulse record at the foot of the ascending limb of the *c* wave, *i.e.*, the beginning of the ejection phase. The first component (1) coincides with isometric contraction and the second group (1') with the beginning of ejection.

The normal phonocardiogram. The *first sound* is usually recorded as a single group of rapid vibrations, which quickly reach a maximum and then rapidly decrease (Fig. 39, 1). If the sounds are captured over the apex, the first sound is the most intense one; *i.e.*, the amplitude of the vibrations is greatest. It lasts 0.10 to 0.17 sec., and the oscillations have a frequency of 25 to 50 per second. The beginning of the first sound coincides exactly with the beginning of ventricular systole, but sometimes this is difficult to detect because the oscillations of the physiologic auricular sound fuse with those of the first sound.

Careful examination of a large number of phonocardiograms of healthy subjects, obtained from the different auscultation areas, shows that in the major-

ity of cases the first sound consists of two definite groups of vibrations separated by a very brief aperiodic interval (Fig. 40, 1 and 1′). The first group corresponds to the isometric contraction, and may therefore be called the isometric component of the first sound; the second group corresponds to the beginning of the ejection phase, and may therefore be called the ejection component. Each one of the groups has two parts, more or less clearly marked. The first vibrations (one or two) of the isometric component have less amplitude, and their frequency is lower. Probably this is due to the modifying influences of the final vibrations caused by auricular systole and to the cardiac impulse on the chest wall at the beginning of ventricular systole. The last part of the ejection component is easily differentiated, probably because of the changes caused in the sound by the acceleration of the blood stream during the maximum ejection phase.[1]

When the separation between the isometric and ejection components is more marked (*e.g.*, owing to respiratory conditions), there is a split first sound. Sometimes the auricular sound can produce a sensation on the ear that simulates a splitting of the first sound.

The *second sound* is recorded as a group of four to six ample vibrations, of frequency slightly greater than that of the first sound (Fig. 39, 2). Its duration is 0.10 to 0.14 sec. If the vibrations have been captured on the upper part of the thorax, those of the second sound are usually the ones with greatest amplitude.

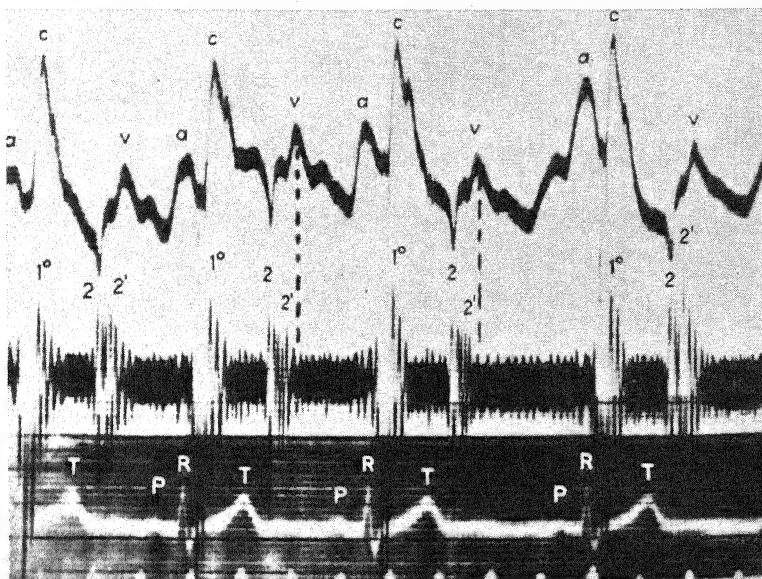



FIG. 41. Reduplication (splitting) of the second sound. Phlebogram, phonocardiogram, electrocardiogram, and time in 0.2 sec. The second sound is made up of two groups of vibrations, 2 and 2′. The second group of vibrations begins before the summit of the *v* wave. Reduplication of the second sound is usually attributed to asynchronous closure of the aortic and pulmonary semilunar valves.

The beginning of the second sound coincides with the bottom of the incisura of the aortic or central arterial pulse. In optical records of the venous pulse, the second sound is always marked by several sharp oscillations or at least a notch on some part of the ascending limb of the *v* wave. Usually, but not always, the second sound coincides with the end of the T wave in the electrocardiogram (Figs. 40 and 41).

Sometimes, even in normal subjects, the second sound appears in the phonocardiogram as two groups of vibrations more or less clearly separated from each other (Fig. 41, 2 and 2′). Each of the components lasts a shorter time than that corresponding to the usual single second sound, and one component follows immediately after the other. This is known as "splitting" or "reduplication" of the second sound. Its main distinctive feature in the phonocardiogram is that the second component of the split sound is registered before the summit of the *v* wave of the phlebogram obtained simultaneously (Fig. 41).

[1] CAEIRO, A., and O. ORÍAS, *Rev. argent. de cardiol.*, 4, 71, 1937.

Splitting of the second sound is usually attributed to asynchronous closure of the semilunar valves of the aorta and pulmonary arteries. Phonocardiograms show that splitting of the second sound is less frequent than it appears to be on auscultation, because the ear easily mistakes the third sound for splitting of the second sound. The only way of arriving at a definite conclusion is to make a simultaneous record of the heart sounds and the phlebogram, or the apex beat.

The *third physiologic heart* sound is recorded more often in phonocardiograms as the sensitivity of the instruments increases. The apex and mesocardiac areas are the most favorable for registering it. In young adults with a thin chest wall it is more frequently registered than in children or middle-aged adults. It is marked in the phonocardiogram by one to six vibrations, or even more (Figs. 39 and 40). Its salient characteristic is the moment in the cardiac cycle at which it occurs. It always coincides with the final portion of the rapid ventricular filling phase, and comes half or two-thirds of the way down the descending limb of the *v* wave of the phlebogram (Figs. 39 and 40) recorded simultaneously.

The interval between the beginnings of the second and third sounds marks the delay with which the ventricle fills after the closure of the semilunar valves, *i.e.*, the duration of the isometric relaxation and rapid filling phases. In normal subjects these phases have an aggregate and very constant duration of 0.11 to 0.14 sec.

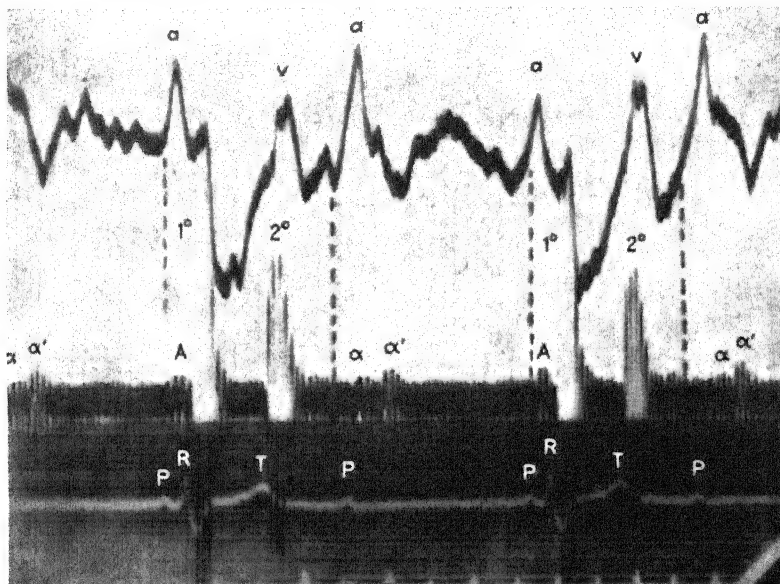



FIG. 42. Phonocardiogram in partial A-V block. Phlebogram, phonocardiogram, electrocardiogram (lead II), and time in 0.2 sec. The phlebogram and ECG show the existence of partial A-V block of the 2:1 type. The phonocardiogram shows two groups of vibrations, α and α', connected with auricular contraction. (Compare with Fig. 72.)

The intensity of the third sound varies in different subjects, and also in the same individual according to the respiratory phase in which it is recorded. The amplitude of its vibrations is usually less than that of the second sound. The duration varies from 0.07 to 0.10 sec.

The *physiologic auricular sound* is not always marked on the phonocardiograms. It appears as a group of vibrations of small amplitude, which precedes the first sound. It begins 0.02 to 0.04 sec. after the beginning of the *a* wave of the phlebogram (Fig. 39). It lasts 0.08 to 0.10 sec. Frequently the auricular sound is not clearly recorded, and care must be taken to recognize it (Fig. 41). In the majority of cases in which auscultation reveals a split first sound, the phonocardiogram shows an auricular sound.

If the heart sounds are recorded with an exploring catheter placed in the esophagus, the auricular sound is always found in the phonocardiogram. It begins earlier and lasts longer than in records obtained from the precordial area (Taquini, 1936).

In cases of complete or partial A-V block, the

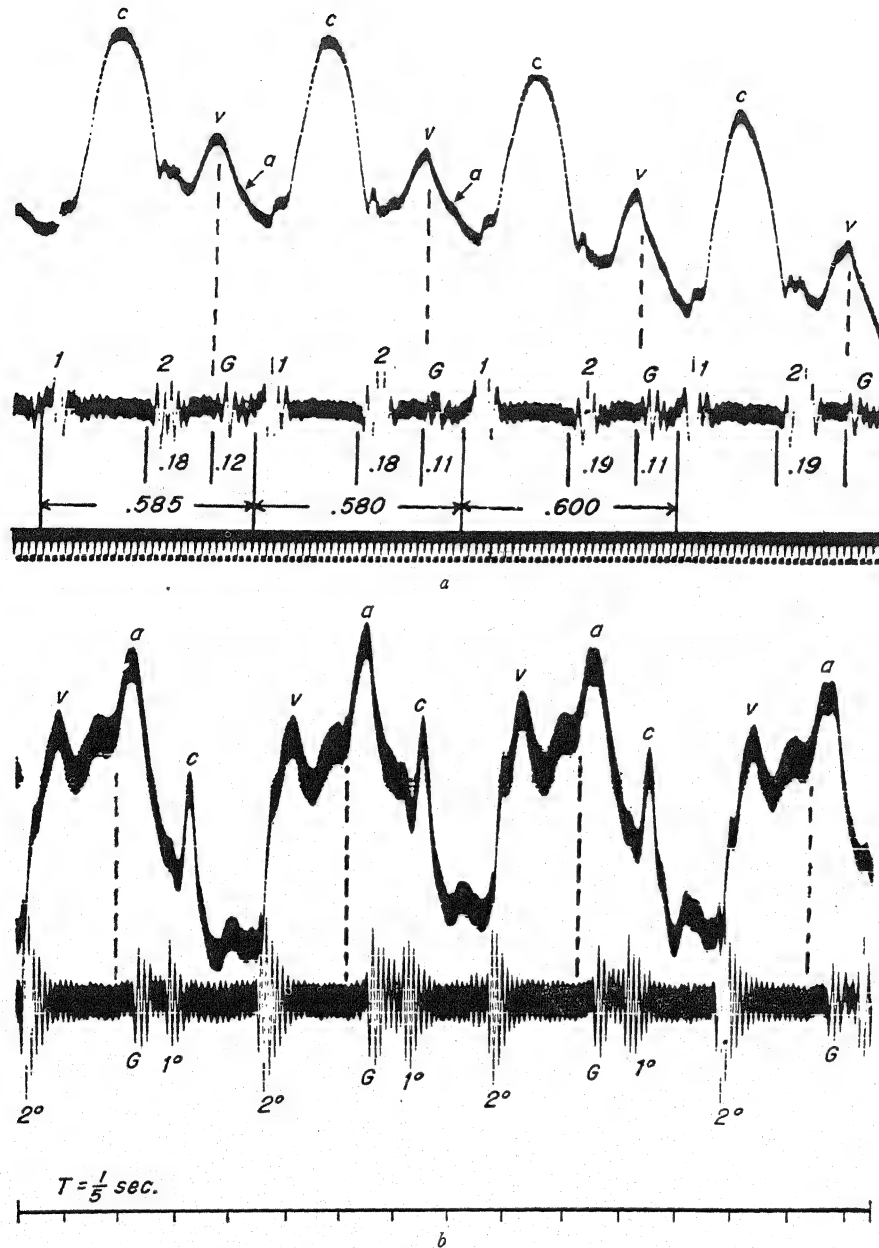



FIG. 43. Gallop rhythms. Above, records of a patient with *rapid-filling* gallop rhythm. Phlebogram, phonocardiogram, and time in 0.2 sec. The *a* wave of the venous pulse appears as a change in gradient in the descending limb of *v*, which is well marked. The *c* wave is very large ("arterialized" phlebogram). The phonocardiogram shows the first (1) and second (2) heart sounds and a third sound (*G*) which is coincident with the descending limb of *v*, and which reaches its maximum intensity before auricular systole (*a*). This rhythm might be mistaken for a presystolic gallop, as it begins nearer the commencement of the following first sound (0.11 to 0.12 sec.) than the end of the preceding second sound (0.18 to 0.19 sec). The simultaneously recorded phlebogram removes all doubt as to its being a rapid-filling gallop rhythm.

Below, phlebogram, phonocardiogram, and time in 0.2 sec. The phonocardiogram shows the first (1°) and second (2°) heart sounds and a third sound *G*. The third sound coincides with the *a* wave in the phlebogram, *i.e.*, auricular systole; it is, therefore, a *presystolic gallop rhythm.* In this case there is also a very slight delay in A-V conduction, but this is not a necessary condition for the existence of this type of gallop rhythm.

auricular sound is registered as two definite groups of vibrations (Fig. 42).

Sometimes extraneous, extracardiac sounds are registered in the phonocardiogram. The different frequency of the vibrations, the irregularity in their occurrence, and the absence of a definite relation to any event in the cardiac cycle easily permit their classification as extraneous sounds.

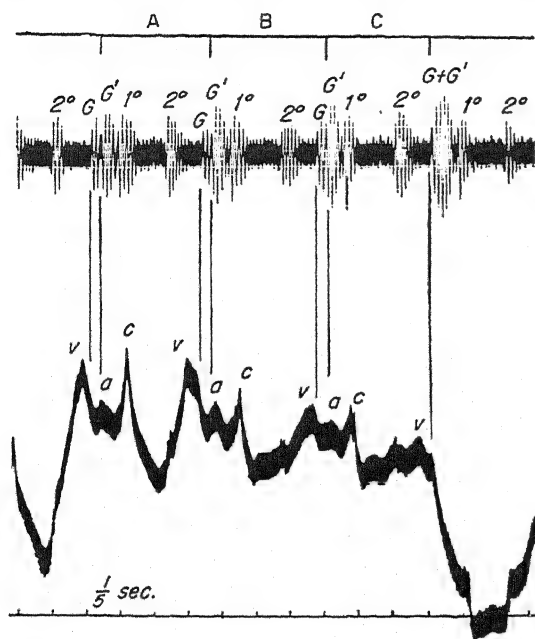



FIG. 44. Records of a patient with complete and incomplete summation gallop rhythms. From above down: duration of each cardiac cycle, phonocardiogram, phlebogram, and time in 0.2 sec. In cycles *A* and *B*, beside the first and second heart sounds, there are two added sounds (*G* and *G'*) very close together, but clearly different from each other. *G* coincides with the descending limb of the *v* wave in the venous pulse (rapid inflow) and *G'* with the *a* wave (auricular systole). Cycle *C* is shorter, because of respiratory arrhythmia, and the end of rapid inflow coincides with the auricular systole; therefore *G* and *G'* have been summated. This is a case of summation gallop rhythm, incomplete in cycles *A* and *B*, and complete in cycle *C*.

THE HEART SOUNDS IN SOME ABNORMAL CONDITIONS

As an introduction to the physiopathology of heart sounds, which has great importance in medical practice, the disturbances produced in a few pathologic conditions will be examined in the following paragraphs.

Gallop rhythms. According to Potain's definition, a gallop rhythm is a triple heart sound made up by the addition of a third component to the two sounds usually heard. The added sound is placed between the two others and is clearly distinguishable from them. The peculiar cadence made Bouillaud give it the name of gallop rhythm.

Phonocardiographic studies have shown that the added sound is always produced at definite moments of the cardiac cycle. Sometimes it coincides with the end of the rapid filling, *i.e.*, with the moment of occurrence of the third physiologic sound (Fig. 43). In other cases it coincides with auricular systole, *i.e.*, with the moment of occurrence of the physiologic auricular sound (Fig. 43). Finally, in the majority of cases it coincides with both the end of rapid filling and the auricular systole, which take place almost at the same time or even simultaneously because of the shortening of diastole as a result of tachycardia. In this last case two different circumstances are summated, each one of which is capable of producing on its own a third sound responsible for the gallop rhythm. Sometimes summation is not complete, and a very brief interval between the two sounds can be seen in the records (Fig. 44, cycles *A* and *B*); at others there is complete summation, and a single sound results which has an intensity greater than that of its components (Fig. 44, cycle *C*).

The added sound in the gallop rhythm is therefore only an exaggeration of normal acoustic phenomena, which occur in the course of cardiac activity and which usually pass unnoticed on auscultation. Clinical experience shows that this exaggeration occurs when there are serious disturbances in the myocardium which cause ventricular hypotonicity (hyperdistensibility). In all cases analyzed so far, the extra sound occurred in diastole, and according to the mechanism of its production, the following types of gallop rhythm can be distinguished: *presystolic gallop*, *end-of-rapid-filling gallop*, and *summation gallop*. Summation can be either complete or incomplete (Figs. 43 and 44).

A systolic gallop rhythm occurs when the added sound is interposed between the first and second normal sounds, during ventricular systole. The added sound has been attributed to the vibrations caused by the sudden distention of a fibrous cord or membrane.

Valvular opening snaps. Normally the opening of the A-V valves is a silent event, but when the valvular orifice is narrowed by pathologic processes such as adhesions or fibrous cords, the opening of the valve at the end of diastolic relaxation is hindered and the

valve becomes suddenly distended, giving rise to a peculiar snapping sound. This occurs in cases of mitral stenosis, a frequent result of rheumatic cardiac disease.

In such cases a third sound appears in the phonocardiogram, the beginning of which coincides with the summit of the *v* wave of the phlebogram simultaneously recorded. The summit of the *v* wave is due to the opening of the A-V valves; so the name "opening snap of the mitral valve" is the appropriate one for this sound (Fig. 45).

A diagram of the different possibilities of normal and abnormal heart sounds in relation to the waves of the venous pulse is given in Fig. 46.

Fig. 45. Opening snap of the mitral valve. Phlebogram, phonocardiogram, electrocardiogram, and time in 0.2 sec. Records of a case of mitral stenosis. The sound *Sn* which follows the second heart sound begins coincidentally with the summit of the *v* wave; therefore it corresponds to the opening of the A-V valves.

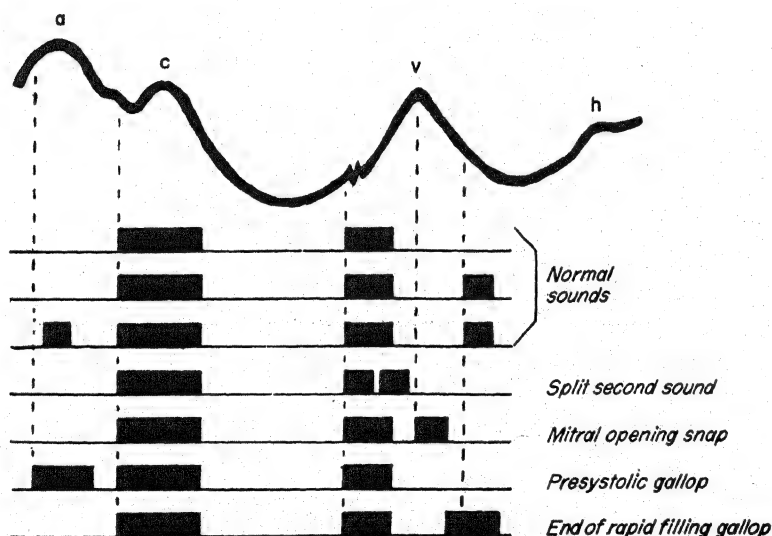



Fig. 46. Diagram of phlebogram and heart sounds, showing the time relations between the waves in the venous pulse and the principal types of normal and pathologic heart sounds.

Cardiac murmurs. When the flow of blood becomes turbulent (with the formation of eddies) within the heart and the velocity of the flow is greater than a certain critical speed, the ear perceives a peculiar sound called a murmur. These conditions occur in the heart when the blood flows through narrowed valvular orifices or flows back through an incompetent valve (valvular stenosis or valvular insufficiency). Murmurs are not always well recorded in the phonocardiogram.

A murmur does not always mean a valvular lesion. If the velocity of flow is increased, as in hyperthyroidism, murmurs may be heard in the absence

of valvular lesions. In cases of severe anemia, murmurs occur because the flow easily becomes turbulent, owing to the low viscosity of the blood.

BIBLIOGRAPHY

BONDI, S., Die Entstehung der Herzgeräusche, *Ergebn. inn. Med. u. Kinderheilk.*, **50**, 308, 1936.

BOYER, N. H., R. W. ECKSTEIN, and C. J. WIGGERS, Normal Heart Sounds, *Am. Heart J.*, **19**, 257, 1940.

CALÓ, A., "Les Bruits du coeur et des vaisseaux," Masson et Cie, Paris, 1950.

LEVINE, S. A., and W. P. HARVEY, "Clinical Auscultation of the Heart," Saunders, Philadelphia, 1949.

MANNHEIMER, E., Calibrated Phonocardiography and Electrocardiography, *Acta paediat.*, **28**, Suppl. II, 1940

ORÍAS, O., and E. BRAUN MENÉNDEZ, "The Heart Sounds in Normal and Pathologic Conditions," Oxford, New York, 1939.

ORÍAS, O., and E. BRAUN MENENDEZ, Der Vorhofton des Herzens, *Ergebn. d. Physiol.*, **43**, 57, 1940.

RAPPAPORT, M. B., and H. B. SPRAGUE, The Graphic Registration of the Normal Heart Sounds, *Am. Heart J.*, **23**, 591, 1942.

SCHMIDT-VOIGT, J., "Herzschalldiagnostik in Klinik und Praxis," Thieme, Stuttgart, 1951.

SCHOLER, H., and M. SPRINGENSCHMID, Bemerkungen zur Phonokardiographie, *Schweiz. med. Wchnschr.*, **80**, 807, 1950.

SCHÜTZ, E., Physiologie des Herztöne, *Ergebn. d. Physiol.*, **35**, 632, 1933.

SEGURA, A. S., Exploración Gráfica de los Ruidos Cardíacos en el Lactante Normal, *Rev. argent. de cardiol.*, **3**, 3, 1936.

TAQUINI, A. C., "Exploración del Corazón por Vía Esofágica," El Ateneo, Buenos Aires, 1936.

TRENDELENBURG, F., "Klänge und Geräusche," Springer, Berlin, 1935.

CHAPTER 15

The Electrical Changes during Cardiac Activity

THE ACTIVITY OF the heart muscle is accompanied by electrical phenomena, similar in nature to those which take place when any tissue becomes active. There is still much to be learned about the intimate nature and the mechanism of these electrical phenomena.

The use of instruments easily handled and sufficiently faithful to record them, however, has greatly advanced our knowledge of this particular aspect of the heart action, in both normal and pathologic conditions. No medical examination of the heart is now considered complete if its electrical changes have not been recorded.

The electrical phenomena of the heart were discovered by Kölliker and Müller in 1856.[1] They observed that when the sciatic nerve of a frog was placed on a beating heart of the same or another animal, at each systole the muscles innervated by the sciatic contracted. They attributed the stimulation of the nerve to an electric current originated by the activity of the heart. This fundamental experiment can be repeated in the dog. The thorax is opened in an anesthetized animal kept alive by artificial respiration, and the phrenic nerve is dissected and cut. The diaphragm on the corresponding side will contract synchronously with the heartbeat whenever the peripheral end of the cut nerve is placed on the heart.

Waller[2] was the first to obtain a graphic record of the action current of the normal heartbeat, by means of Lipmann's capillary electrometer, but the first important advance in the study of the electrical phenomena of the heart took place in 1903, when Einthoven, then professor of physiology of Leyden, adapted the string galvanometer for the registration of the action currents of the heart and subsequently developed the fundamental theory of electrocardiography. The clinical and experimental studies of Lewis, Wiggers, F. N. Wilson, and others have contributed to enlarge and consolidate the basic knowledge necessary to interpret the electrical variations inherent in cardiac activity.

[1] KÖLLIKER, R. A., and J. MÜLLER, *Verhandl. d. physik-med. ges. Würzburg*, **6**, 530, 1856.
[2] WALLER, A., *J. Physiol.*, **8**, 231, 1887.

The action current of the heart is transmitted to the neighboring structures. It may therefore be recorded by connecting the galvanometer to convenient points of the body surface without even damaging the skin. When the action current is led off directly from the heart, the record is called an "electrogram" (Samojloff),[1] the term "electrocardiogram" (Einthoven) being used when the record is obtained indirectly by leads placed on the skin.

RECORDING APPARATUS

The changes in electrical potential originated by the heart action are of very small magnitude and occur in relatively fast succession. The instruments for recording them must therefore possess great sensitivity and rapidity of response (high natural frequency). The *string galvanometer*, as developed by Einthoven (1903), meets these requirements. It is based on the following principle: when an electric current flows through a conductor placed at right angles to a magnetic field, the conductor moves in a direction at right angles both to the magnetic field and to the direction of the current (Fig. 47).

Several models of electrocardiographs are manu-

[1] SAMOJLOFF, A., *Pflüger's Arch. f. d. ges. Physiol.*, **135**, 422, 1910.

factured based on the principle of Einthoven's string galvanometer, from which they differ only in details.

String galvanometers are not the only instruments used to record the action currents of the heart. Amplification of very weak currents by thermionic valves has made possible the use of oscillographs for

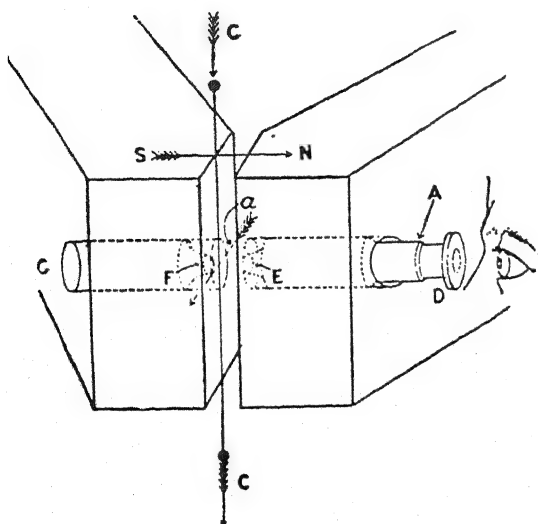



FIG. 47. Diagram of a string galvanometer. *S* and *N*, poles of an electromagnet; *C-C*, the string of the galvanometer; *C-F*, system of lenses; *E-D*, microscope; *A*, arrow indicating the direction in which the string moves when conducting a descending current. (*After T. Lewis.*)

this purpose. There are several types of oscillograph. Instruments are now made in which photographic recording has been replaced by a direct method. A stylet traces the curve in ink, or it is warmed so that it leaves a trace on special waxed paper. These instruments have the advantage that the record does not have to be developed but can be read immediately. They are quite as efficient as optical records.[1]

THE ELECTROCARDIOGRAM

An electrocardiogram is a graphic record of the action current of the heart led off from the body surface. There is an infinite number of points on the body surface from which this current can be led off and recorded. In order to obtain comparable results there is general agreement to place the electrodes on certain conventional places, described further on in the paragraph on the electrocardiographic leads. The tracings differ according to the lead used, but they all show certain features in common, so that a general description can be given.

The normal electrocardiogram shows at each cardiac cycle, in all the usual leads, three positive deflections or waves (above the basal or isoelectric line) and two negative ones (below the basal line) (Fig. 48). They are named with letters as proposed by Einthoven. The first deflection is positive, and corresponds to the spread of excitation in the auricles; it is known as the P wave (presystolic). The subsequent deflections are called by the letters following P in the alphabet: waves Q, R, S, and T. Exceptionally there is a sixth deflection U.

P, R, and T are positive waves; Q and S are negative. The amplitude of Q and S waves varies considerably, and in some leads they may be scarcely marked or completely absent.

Time intervals can be recorded in several ways. A system commonly used consists in a rotating motor of constant speed. A wheel attached to the motor carries five spokes at equal distances from each other, one of which is wider than the rest. The wheel rotates five times per second, therefore each rotation takes ⅕ (0.20) sec.; every 0.04 sec., one of the spokes interrupts the light for a very short time; the wider spoke will interrupt it for a slightly longer time. A fine vertical line will therefore appear on the developed record every 0.04 sec., and every fifth line will be a little wider than the others.

It has been conventionally decided to adjust the tension of the string in the string galvanometer, or the amplification power in the oscillographs, so that a current of 1 mv. causes a deflection of 1 cm. on the record. The manifest value of cardiac potentials can thus be judged from the amplitude of the deflections.

Description of the waves. The P wave is usually low and rounded, appearing as a wide tracing on the record, because of the relatively low speed at which it develops. It can reach 2 mm. (= 0.2 mv.) in height and lasts about 0.1 sec. (Fig. 48).

The anomalies in the shape, situation, direction, number, etc., of the P waves give important information on the functioning of the auricles. Nevertheless it is necessary to bear in mind that the P wave can vary considerably in size and contour in one or in several leads without this variation having an abnormal significance (Wiggers). Neither is there any relation be-

[1] HUNZICKER, W. J., and H. D. LEVINE, *Am. J. M. Sc.*, **218**, 37, 1949.

tween the size of the P wave and the strength of auricular contraction.

The waves Q, R, S, and T are related to ventricular activity. It is usual to consider them as forming two complexes: an initial, rapid complex (QRS), and a final slow complex (T). They are separated by a short isoelectric interval. The U wave first observed by Einthoven, and later by Hering and by Lewis, seldom appears; when visible it is more marked in lead II (right arm–left leg). Its significance is unknown. Artefacts due to an exceedingly slack string (owing to a high resistance in the subject's circuit) may be taken for U waves.

The Q wave is negative; it is never very prominent in normal subjects in any of the usual leads. It may be absent without having any pathologic significance.

The R wave is positive; it is usually the highest deflection, especially in lead II. The amplitude of R depends on the direction of the electrical axis of the heart (see further on) during the particular moment at which the initial ventricular complex is being recorded. In normal subjects it is 7 to 17 mm. (= 0.7 to 1.7 mv.).

The S wave is negative, varying in depth according to the subject and the lead, depending on the direction of the electrical axis of the heart.

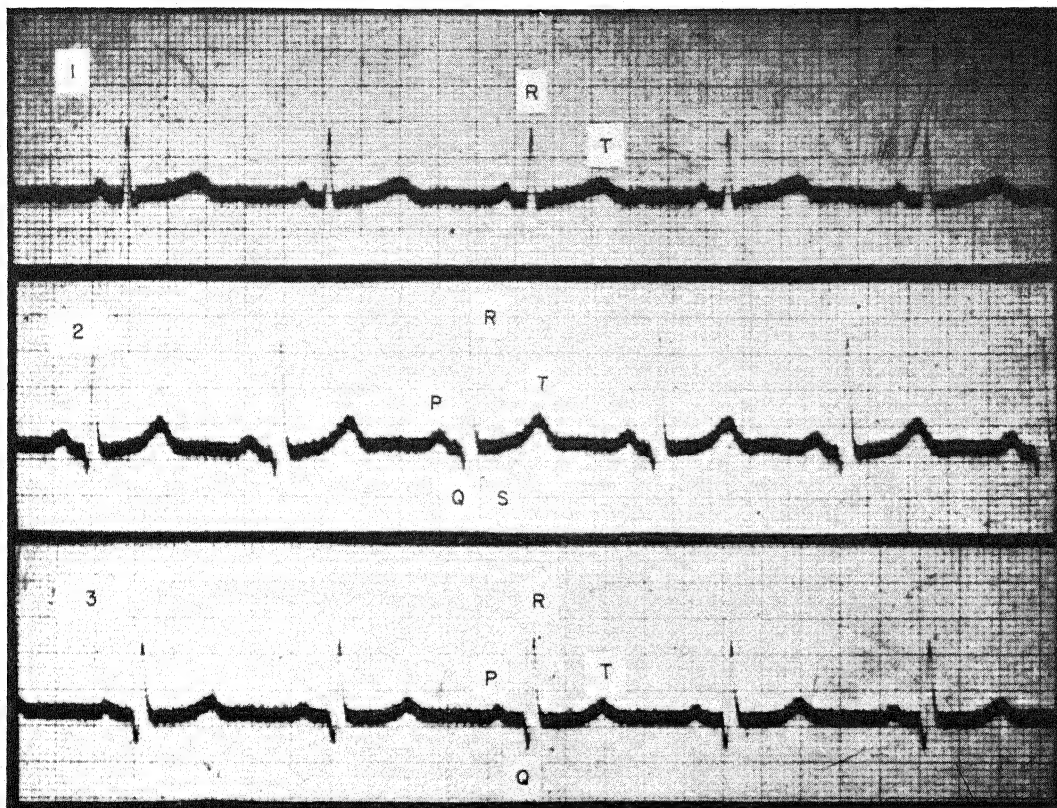



FIG. 48. Normal human electrocardiogram, leads I, II, and III. PR (or PQ) interval, 0.16 sec.; QRS interval, 0.08 sec.; ST segment, isoelectric.

The tracing of the QRS complex in normal subjects is in thin, straight lines, of uniform width, and without slurring, splintering, or notches. These abnormalities denote slight variations in intraventricular conduction and are important when they appear in records taken in more than one lead. They are frequently observed, even in healthy subjects, in lead III (left arm–left leg). In this lead the initial ventricular complex is so irregular in some cases that the three waves cannot be distinguished; the complex is then said to be of vibratory type. When there are no other

anomalies, this has no pathologic significance. Figure 49 shows a record with splintering of the ascending limb of R and thickening of the summit of S.

The T wave in normal human subjects is always positive in leads I and II; but it is sometimes negative in lead III. T is a relatively slow and rounded wave; its height in lead II is approximately one-quarter to two-fifths the height of R. Anomalies in the T wave depict myocardial damage (infectious, toxic, nutritional).

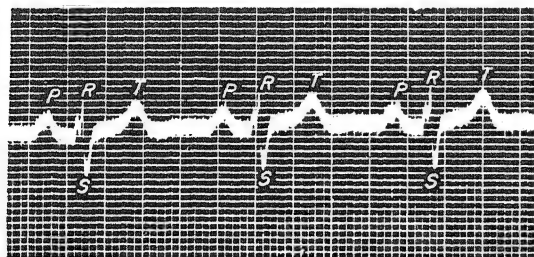



Fig. 49. "Splintered" and "thickened" ventricular complex. The R wave does not have the straight, thin lines of the normal ECG, but has its ascending limb "splintered." The deep S wave has a thickened summit.

Time relations. Certain time relations in the electrocardiogram are of great importance. The interval between the beginning of P and the beginning of Q (R, when Q is missing), known as the PQ, or PR interval, lasts from 0.13 to 0.18 sec.; its variations indicate abnormalities in the time relations between the auricular and ventricular contractions. An interval of more than 0.18 sec. denotes a delay in conduction from the auricles to the ventricles—a first degree of auriculoventricular block.

The QRS complex in normal subjects never lasts more than 0.10 sec.; its duration is usually from 0.06 to 0.08 sec. The duration of QRS is prolonged in cases in which excitation spreads by abnormal routes and one ventricle is stimulated before the other; this occurs in cases of ventricular extrasystoles, or when there is a block in one of the branches of the His bundle (bundle-branch block). In these cases T is usually enlarged and of opposite direction to the QRS complex (Fig. 50).

The total duration of QRST—known as the *electrical systole*—is approximately the same as that of the mechanical systole, but it cannot be used to determine the duration of the contraction of the heart.

The interval between the initial and final ventricular complexes, known as the ST interval, is particularly significant. In normal conditions its duration varies with the shape and size of T, but it is always isoelectric, *i.e.*, the record should be at the same level as during diastole, or not more than 1 mm. above or below this level. A greater deviation indicates anoxia in the myocardium (asphyxia, coronary occlusion, compression of the coronary arteries by pericardial effusion).

Mechanical events and the electrocardiogram. The deflections of the electrocardiogram precede all mechanical events in the activity of the heart to which they are connected. Figure 51 shows that the beginning of the systolic rise in intraventricular pressure occurs when the ascending limb of the R wave of the simultaneously recorded electrocardiogram (lead II) has been completed. The increase in intraauricular pressure also begins when the P wave of the electrocardiogram has reached its summit. Usually the beginning of the electrical deflection precedes the beginning of the mechanical event by 0.01 to 0.03 sec.

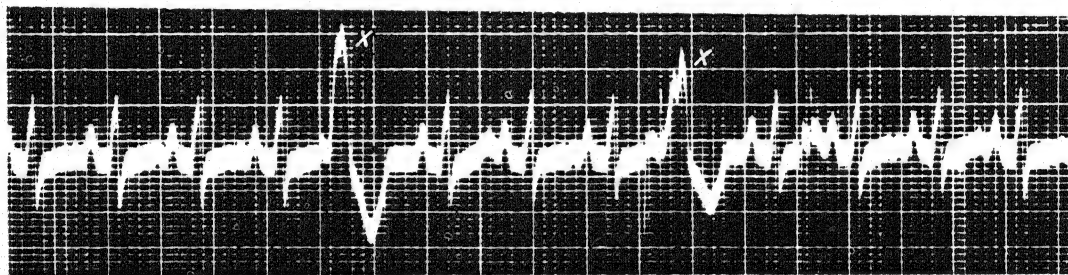



Fig. 50. Ventricular premature beats (X) provoked by direct electrical stimulation of the right ventricle in the dog. The premature beats appear in advance of the normal beat, the QRS complex is lengthened, and there is a large T wave in the opposite direction to the initial complex (lead II).

There is a certain degree of coincidence, by no means absolute, between the end of the T wave and the end of ventricular systole.

There is no relation whatsoever between the strength of ventricular systole and the amplitude of the electrical deflections. A vigorous systole may be accompanied by a small electrical deflection and vice versa.

Before accepting any of the current hypotheses, since none gives a complete explanation, it is best simply to describe these facts, even though they may appear to be disconnected, until further observations permit an integrated and satisfactory interpretation.

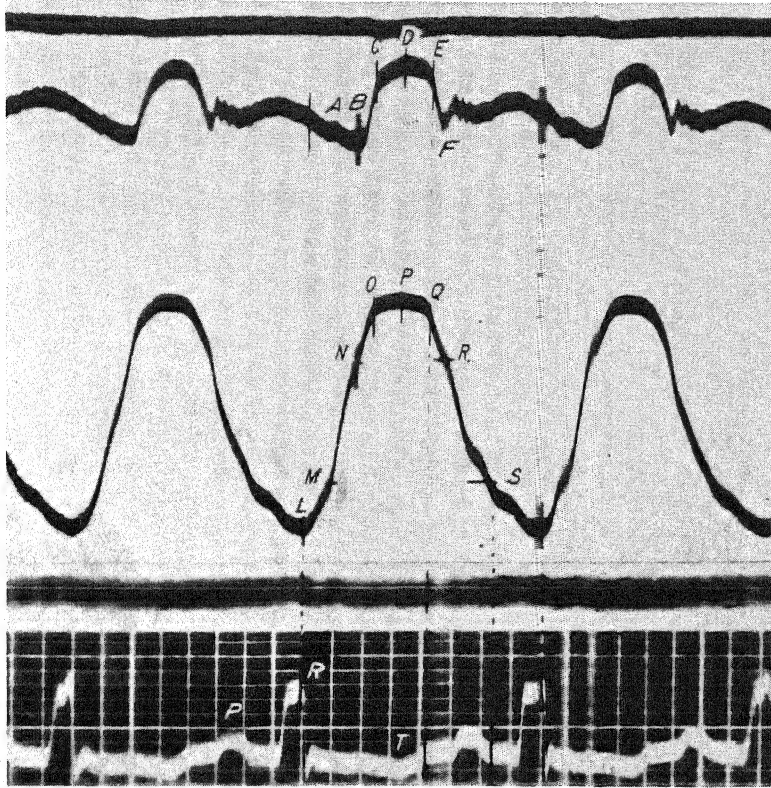



FIG. 51. Simultaneous records of mechanical changes (pressure pulses) and electrocardiogram in the dog. Aortic pressure, left intraventricular pressure, ECG (lead II). Time in 0.04 sec. The T wave is frequently negative in the dog in lead II, as in the case recorded. The first sign of ventricular activity in the intraventricular pressure pulse (*L*) corresponds to the summit of the R wave in the ECG. The end of ventricular systole in the pressure curve (*Q*) corresponds to the end of the T wave in the ECG, but this is not always so. Mechanical changes in the auricle begin at *S*, when the P wave of the ECG has ended.

THE MEANING OF ELECTROCARDIOGRAPHIC WAVES

There is as yet no unanimous agreement regarding the exact meaning of electrocardiographic deflections. There is not even agreement as to the nature of the fundamental electrical process. There are, however, several well-known but more or less isolated facts, which undoubtedly represent only partial aspects of the problem but are important for its complete solution.

Electrical variations in a single isolated myocardial fiber. When a microelectrode, connected with an adequate recording system, is inserted into a single myocardial fiber of the ventricle in the beating heart of a frog, a series of well-marked electric deflections are registered.[1] If the electrode is placed simply in contact with the outside surface of the fiber without penetrating into it, a basal isoelectric line is registered. If the needle punctures the membrane and enters

[1] WOODBURY, L. A., H. H. HECHT, and A. R. CHRISTOPHERSON, *Am. J. Physiol.*, **164**, 307, 1951; WOODBURY, L. A., and H. H. HECHT, *Circulation*, **6**, 172, 1952.

into the fiber during diastole, a sudden electrical potential appears and the galvanometer shows the passage of a direct current with the negative pole inside the fiber. This is the *membrane resting potential*. When the fiber enters into systole, this potential rapidly falls to zero; it then reverses

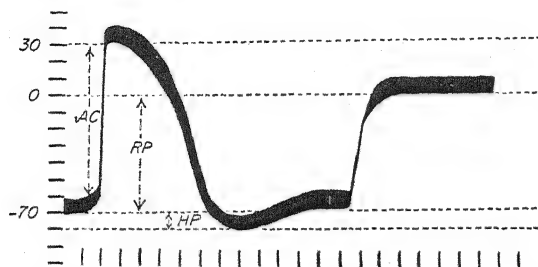



FIG. 52. Membrane potentials of cardiac fibers. *RP*, resting potential; *AC*, action current; *HP*, hyperpolarization following action current. Ordinate, millivolts; abscissa, time in 0.1 sec. (*After Woodbury, L. A., H. H. Hecht, and A. R. Christopherson, Am. J. Physiol., vol. 164, p. 307, 1951.*)

as the inside of the fiber becomes positive with respect to the outside (Fig. 52). This inversion of the intracellular potential is known as depolarization, and it corresponds to the activation of the myocardial fiber. A period of gradual *repolarization* or recovery follows, during which the membrane potential slowly returns to the resting state. This sequence of depolarization and repolarization is known as the "membrane action potential," to distinguish it from action potentials registered on the surface of the cell. If the microelectrode is withdrawn from the cell, whatever voltage is present disappears and the isoelectric line is again recorded. The membrane action potential lasts as long as the action potential of the whole fragment of myocardium or the heart of which the fiber forms part. The average membrane resting potential of several observations was found to be −64.5 mv., and the peak of the membrane action potential was +12.7 mv.

Electrical deflections in myocardial fragments. Since Frédéricq's[1] original observations, it is well known that small fragments of myocardium, even fragments taken from the auricles, show action potentials with a rapid and a slow component similar to the QRS and T components of the electrocardiogram respectively. Records obtained by means of bipolar electrodes placed on the heart or on a fragment of myocardium, even though they are separated by only a few millimeters and lead off action potentials arising in the small mass of underlying tissue, also show a first rapid component similar to QRS followed by a slow component similar to T.

Records taken from myocardial fragments of various shapes and from limited areas of the heart (by means of bipolar electrodes placed near each other) show that the rapid component, similar to QRS, is inverted when excitation spreads from the opposite direction to that taken in a first record. On the other hand, the slow component, similar to T, remains unaltered whatever the direction of the spread of excitation.[1] This fact is important for the interpretation of the significance of electrocardiographic waves, since it suggests that the rapid and slow components are due to two different processes, the former depending on the direction of the spread of the excitation and the latter not (dual origin of waves).

Electrocardiograms of univentricular hearts. The electrocardiograms of hearts with a single ventricle (fishes, amphibians, and some reptiles) are not very different from the electrocardiograms of hearts with two ventricles (other reptiles, birds, mammals). In all of them there is a P wave, corresponding to auricular systole, and two components, QRS and T, corresponding to the ventricular systole. In amphibians usually another small wave, corresponding to the arterial (aortic) bulb, is recorded. The QRS and T components are very similar to the homonymous deflections of mammalian hearts,[2] although slower.

Current interpretations. It is well to keep in mind that the usual interpretations of the intimate processes which give rise to the ECG are merely hypothetical, and that none of them can be considered as completely and definitely established.

According to the classic idea of the fundamental electrical phenomenon, active parts of the myocardium become electronegative with respect to inactive parts (the relative-negativity hypothesis). More recent studies, however, seem to show that a double electrical variation takes place in each active unit, *i.e.*, a negative and a

[1] FRÉDÉRICQ, H., *Arch. internat. de physiol.*, **11**, 234, 1911; **12**, 66, 1912.

[1] AMUCHASTEGUI, S. R., O. ORÍAS, and A. S. SEGURA, *Rev. Soc. argent. de biol.*, **18**, 138, 1948.

[2] LEPESCHKIN, E., "Modern Electrocardiography," Williams & Wilkins, Baltimore, 1951.

positive potential, not simply a relative negativity (the doublet or dipole hypothesis).

The ventricular complex (QRS and T) is due, according to many workers, to a band of dipoles (Fig. 53*A*) which moves along the heart as the process of excitation advances.[1] Others[2] interpret it as due to the algebraic summation of two asynchronic monophasic currents (interference theory) (Fig. 53*B*). On the other hand, Thomas Lewis, interpreting carefully taken records of the spread of excitation through the heart, concluded that the QRS component of the ventricular mammalian ECG is the result of the algebraic summation of electrical variations in each ventricle.

Much has still to be learned about the intimate processes which give rise to electrical phenomena in the heart and about why the ECG shows the deflections that are in fact recorded. Clinical observations and experimental studies have, however, given enough facts to permit the diagnosis and interpretation of many typical functional and organic disturbances in cardiac functions.

THE ELECTRICAL AXIS OF THE HEART

Definition and general idea. A vector symbolizes a magnitude with a definite direction. It is represented by an arrow, the length of which gives the strength and the point of which gives the direction of the force. Physical forces are represented by vectors. The electromotive force developed by the heart at any moment can be symbolized by a vector, figured by an arrow. This vector is called the *electrical axis of the heart*. The length of the arrow indicates the voltage or the *manifest potential* resulting from the electromotive forces developing at any given moment, and the direction of the vector indicates the sense in which the current flows.

The absolute value, *i.e.*, the total electromotive force developed by cardiac activity, cannot be determined by the ECG as it is usually recorded. The manifest potential difference is only a fraction of the actual potential difference, but it changes in direct ratio with the latter and can be established by the amplitude of the deflections recorded in two or more leads.

Fig. 53. Current interpretations of the ECG. *A*, galvanometric record (below) as the dipole (above) approaches, lies under, and recedes from the exploring electrode P. *B*, resultant electrograms of summation of two monophasic potentials 1 2 and 1′ 2′. *C*, left, theoretic curve of a record corresponding to the situation illustrated in *A*; right, tracing of a record obtained experimentally with an exploratory electrode. (*A and C, after Hecht, H. H., and L. A. Woodbury, Circulation, vol. 2, p. 37, 1950. B, after Hoff in Fulton, "Howell's Textbook of Physiology," 15th ed., Saunders, Philadelphia, 1946.*)

The electrical condition of the heart during systole can be defined as a complex system of dipoles arising more or less simultaneously in the heart, which is surrounded, without any interposing isolating layer, by conducing tissues. Physically this is the equivalent of a system of multiple charges arising in a sphere submerged in a weak conducting medium.[1] These charges can be simplified at any given moment into a resultant dipole, and can be represented by a vector in space.

Einthoven's[2] mathematical and experimental

[1] CRAIB, W. H. (see bibliography at the end of the chapter); WILSON, F. N., *et al.* (see bibliography at the end of the chapter); BAYLEY, R. H., *Proc. Soc. Exper. Biol. & Med.*, **42**, 699, 1939; ASHMAN, R., W. S. WILDE, and C. E. DRAWE, *Am. J. Physiol.*, **128**, 547, 1940; HECHT, H. H., and L A. WOODBURY, *Circulation*, **2**, 37, 1950.

[2] SCHÜTZ, E., *Ergebn. d. Physiol.*, **38**, 493, 1936; SCHÜTZ, E., K. E. ROTHSCHUH, and C. E. MERING, *Klin. Wchnschr.*, **19**, 9, 1940; HOFF, H. E., L. H. NAHUN, and B. KISCH, *Am. J. Physiol.*, **131**, 687, 1941; SCHAEFER, H., "Elektrophysiologie," vol. 2, Deuticke, Vienna, 1942; ROTHSCHUH, K. E., and E. SCHÜTZ, *Klin. Wchnschr.*, **24**, 673, 1947

[1] FATTORUSSO, V. M., THAON, and J. TILMANT, *Acta cardiol.*, **4**, 464, 1949.

[2] EINTHOVEN, W., G. FAHR, and A. DE WAART, *Arch. f. d. ges. Physiol.*, **150**, 275, 1913; English translation in *Am. Heart J.*, **40**, 163, 1950.

analysis showed that it is possible to establish the electrical axis of the heart by leading off currents from different angles, because the amplitude of the deflections in the ECG varies with the inclination of the derived current. When the derived current (the line joining the two electrodes) is parallel to the electrical axis, the deflections are maximal.

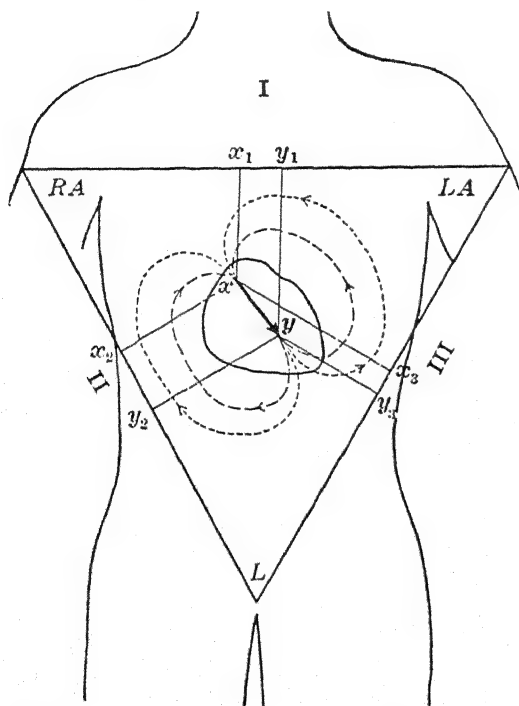



FIG. 54. The electrical axis of the heart. *RA*, right arm; *LA*, left arm; *I*, lead I; *II*, lead II; *III*, lead III. The deflections caused by the action current of the heart in each lead will correspond in size to the projection of the arrow on each side of the triangle (x_1y_1; x_2y_2; x_3y_3). (*From Wiggers, C. J., "Physiology in Health and Disease," 4th ed., Lea & Febiger, Philadelphia, 1944.*)

During the electrical systole the electrical axis is continuously changing in magnitude and direction. In electrocardiographic studies the following aspects are usually considered: (*a*) the instantaneous electrical axis; (*b*) the mean electrical axis; (*c*) the ventricular gradient; and (*d*) the complete evolution of the cardiac vector during the whole cycle (vectorcardiogram).

Instantaneous electrical axis. The instantaneous electrical axis, or instantaneous cardiac vector, is the vector representing the resultant of all the electromotive forces of the heart at a given moment. Its projection on a frontal plane can be calculated from the ECG obtained by the three limb leads. Einthoven represented the three standard leads by the three sides of an equilateral triangle with the heart in the center. The electromotive forces arising in the heart at any given moment can be represented by the electromotive force of the equivalent dipole, the magnitude and direction of which can be figured by a vector symbolized by $\widehat{E}$. On each side of the triangle the amplitude of the deflection of the corresponding lead is laid on to the right or the left of the zero point according to the sign of the deflection. Perpendicular lines from these points are traced toward the center of the triangle (Fig. 54). The instantaneous axis can be calculated at any moment of cardiac activity by projecting the amplitude of the deflections in two or more leads at that moment.

Mean electrical axis. The mean electrical axis represents the average of all the instantaneous vectors corresponding to the QRS groups, or to T or P waves. Usually the mean electrical axis of QRS and T are the ones calculated, establishing the size, direction, and sign of the vectorial quantities. The electrical axis for QRS ($\hat{A}_{QRS}$) is known as the mean electrical axis of depolarization, and $\hat{A}_T$ as the mean electrical axis of repolarization, in agreement with the hypothesis that connects these deflections with the respective electrical processes.

The mean electrical axis for QRS is calculated by determining the area of the positive waves, *i.e.*, those above the isoelectric level (R wave), and subtracting the area of the negative waves, *i.e.*, those below the isoelectric level (Q and S). This *net area* of QRS is determined for the deflections of at last two leads, *e.g.*, lead I and lead II, and the figures obtained are projected on the corresponding side of the Einthoven triangle. The point where the perpendiculars meet inside the triangle gives the vector or mean electrical axis for $\hat{A}_{QRS}$ projected in the frontal plane (Figs. 55 and 56).

The electrical axis of the T wave $\hat{A}_T$ is established in a similar way by projecting on the corresponding sides of the triangle the net area of T in at least two leads. If T is biphasic, the negative area is subtracted from the positive area.

The surface covered by the deflections can be determined by means of a planimeter, after amplification of the record, if necessary. For practical purposes

sufficiently accurate results can be obtained by considering the deflections as triangles and calculating the surface by multiplying the base by half the height (amplitude).[1] The areas thus obtained and the resulting vector represent microvolts per second. Electrocardiographic paper or film commonly in use has the

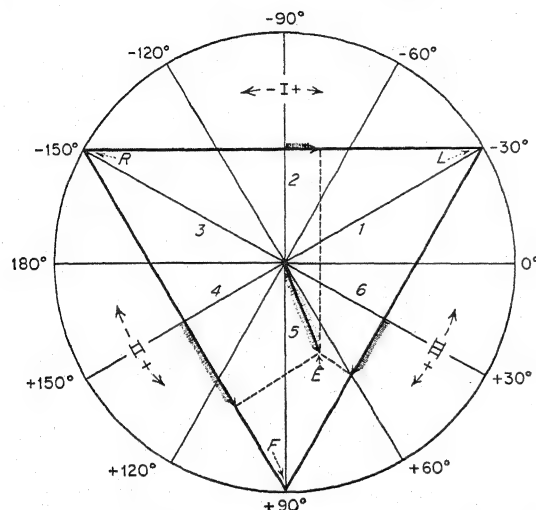



FIG. 55. Einthoven's triangle and triaxial system for determining the electrical axis of the heart.

vertical lines separated by 0.04-sec. intervals and the horizontal lines at 1-mm. intervals, so that each small square corresponds to 4μ v/sec. This is known as one Ashman unit (AU), and the area of the deflections is usually given in Ashman units when calculating the mean electrical axis. Einthoven's triangle can be substituted by a triaxial system, derived from the triangle (Figs. 55 and 56).

The direction of the manifest mean axis QRS ($\hat{A}_{QRS}$) is defined by Einthoven's α angle (see next paragraph), and its magnitude in microvolts per second, *i.e.*, in Ashman units.

Variations in the electrical axis of the heart. The direction of the electrical axis is not strictly dependent on the direction of the anatomical axis of the heart. The electrical axis is also determined by the relative predominance of the ventricular masses and the path of the spread of the process of excitation along the heart.

The angle formed by $\hat{A}_{QRS}$ (mean axis of QRS) or $\hat{E}_R$ or $\hat{E}_S$ (instantaneous axis for the peaks of R or S, according to which is the predominant deflection) and the horizontal line is called the α angle. It has been conventionally established that α values begin on the left of the subject in the axis corresponding to 3 o'clock. They are considered to be positive when falling below this horizontal line, and negative above it. Bayley[1] has proposed the division of the angles into sextants of 60° each, numbering them counterclockwise beginning on the left of the patient (Fig. 55).

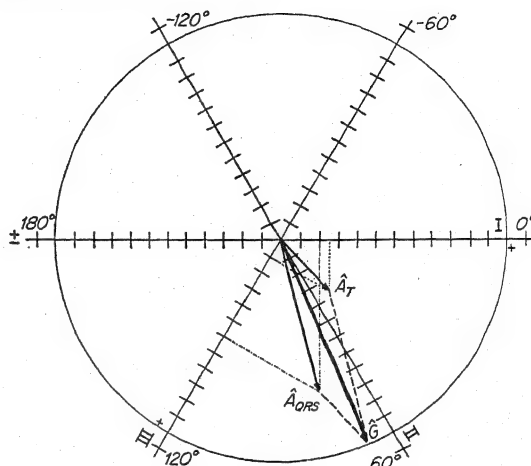



FIG. 56. Ventricular gradient. Vector $\hat{G}$ (the ventricular gradient) is obtained by algebraic summation of vectors $\hat{A}_{QRS}$ and $\hat{A}_T$. The divisions on the radii correspond to Ashman units.

The direction of the normal mean QRS axis varies considerably with the age of the subject. In infants less than 6 months old the axis is markedly deviated to the right (+130°). Between the first and fifth year the axis shifts toward the left; the average for these years is +52°. At puberty it again shifts to the right (+67°), and in the adult it moves to approximately +60°.

In abnormal conditions (ventricular hypertrophy, bundle-branch block, ventricular ectopic beats) the axis may show most variable directions. Table 16 gives the values of α and the net areas of QRS (considered in block as positive or negative) in leads I, II, and III, and the unipolar limb leads, for the different positions of the electrical axis.

Ventricular gradient. In normal conditions the quantity of electricity evolved in the process of excitation (depolarization phase) should be the same, though of opposite sign, as that evolved during the process of recovery (repolarization phase). The area of QRS, if these deflections

[1] ASHMAN, R., and R. H. BAYLEY, *Am. Heart J.*, **25**, 16, 1943.

[1] BAYLEY, R. H., *Am. Heart J.*, **26**, 769, 1943.

are admitted as the electrocardiographic sign of depolarization, should be the same as the area of T, if this wave is considered as the sign of repolarization. The sum of the two areas should be zero, since they are of opposite sign, but it is found that this is not so; therefore, if the hypothesis is still to be considered valid, it must be admitted that the process of repolarization differs from the process of depolarization by some sort of deviation. This break, or drop, or gradient in the physiological uniformity of the heart muscle can be determined quantitatively, according to Wilson and his associates,[1] by the algebraic summation of the net areas of QRS and T. The number thus obtained is called the ventricular gradient.

Independently of the validity of the hypothesis on which it is founded, determination of the ventricular gradient has become a common procedure in electrocardiographic analysis which may allow important deductions.[1]

The ventricular gradient is, therefore, the sum of the vector quantities of QRS and T, *i.e.*, $\hat{A}_{QRS} + A_T$, determined by the procedure explained in the preceding section.[2] The resultant (or sum) of the two vector quantities is equal, in magnitude and direction, to the diagonal of the parallelogram of which they form two sides. The ventricular gradient $\hat{G}$ is determined by first establishing the mean axes of QRS and T, and then adding the vectors that represent them (Fig. 56). The frontal projection of the gradient is thus obtained. It should not be overlooked, however, that the ventricular gradient, like the vectors from which it derives,

Table 16. Classification of the Mean Electrical Axis of QRS According to Its Orientation

Angle of $\hat{A}_{QRS}$	*Net area of QRS in leads*						*Classification*
	I	II	III	*aVR*	*aVL*	*aVF*	
−120°	−	− −	−	+	0	−	Complete inversion of axis
							Inversion to the right
−150°	−	−	0	+ +	−	−	
							Extreme deviation to the right
−180°	− −	−	+	+	−	0	
							Marked deviation to the right
+150°	−	0	+	+	− −	+	
							Moderate deviation to the right
+120°	−	+	+ +	0	−	+	
							Slight deviation to the right
+ 90°	−	+	+	−	−	+ +	
							Tendency to deviation to the right
+ 60°	+	+ +	+	−	0	+	Normal axis
							Tendency to deviation to the left
+ 30°	+	+	0	− −	+	+	
							Slight deviation to the left
0°	+ +	+	−	−	+	0	
							Moderate deviation to the left
− 30°	+	0	−	−	+ +	−	
							Marked deviation to the left
− 60°	+	−	− −	0	+	−	
							Extreme deviation to the left
− 90°	0	−	−	+	+	− −	
							Inversion to the left
−120°	−	− −	−	+	0	−	Complete inversion of the axis

[1] Wilson, F. N., A. G. MacLeod, and F. D. Johnston, *Am. Heart J.*, **10**, 46, 1934; *Tr. A. Am. Physicians*, **46**, 29, 1941.

[1] Ashman, R., and E. Byer, *Am. Heart J.*, **25**, 36, 1943; Bayley, R. H., *Am. Heart J.*, **26**, 769, 1943; Bayley, R. H., J. S. La Due, and D. J. York, *Am. Heart J.*, **27**, 164 and 657, 1944; Bayley, R. H., and J. S. La Due, *Am. Heart J.*, **28**, 54 and 233, 1944.

[2] Ashman and Bayley, *op. cit.*, p. 16.

should really be considered in space, not simply in one plane projection.

The average manifest potential of Ĝ (size in frontal projection) is 13.0 units (1 unit = 4 μv/sec.), slightly more in males than in females, with a maximum around 23.0 and a minimum around 2.5 units. The direction varies normally between 0° and +90°. The ventricular gradient varies with the heart rate; it has a tendency to decrease as the rate increases.

Ventricular-gradient determinations are useful for interpreting the significance of T changes. If a variation in T is due simply to a modification in the anatomical position of the heart, the ventricular gradient will remain normal. If the change in T is due to a pathologic process in the myocardium, the ventricular gradient will be abnormal in both size and direction. Coronary occlusion causes typical alterations in the ventricular gradient (see Ashman, Bayley, etc.).

Vectorcardiography and Vectorcardiogram

The vectorcardiogram is the record showing the evolution in time of the representative vector of the resultant of the electromotive forces arising during the whole cardiac cycle. The vectors of each moment are determined, and their apexes are joined by a continuous line (Fig. 57). The object of vectorcardiography is to obtain vectorcardiograms in order to analyze the electrical activity of the heart.

Einthoven, Fahr, and De Waart[1] determined the instantaneous vectors in leads I, II, and III for 10 moments at equal intervals during the recording of the QRS component. Mann[2] joined the apexes of these vectors and obtained graphs which were called vectorcardiograms. The cathode-ray oscillograph (see Chapter 66) is particularly appropriate for this purpose as it gives directly and automatically the tracing which includes all the resultant vectors, moment by moment, throughout the cardiac cycle, projected on any desired plane (plane vectorcardiography), so that the vectors can be considered also in space (spatial vectorcardiography or stereovectorcardiography). Schellong[3] and Wilson and his associates[4] were the first to use the cathode-ray oscillograph for obtaining plane vectorcardiograms. Later the method has been used to analyze the vectors in space.[1]

[1] Einthoven, Fahr, and De Waart, *loc. cit.*
[2] Mann, H., *Arch. Int. Med.*, **25**, 283, 1920.
[3] Schellong, F., *Verhandl. d. deutsch. Gesellsch. f. inn. Med.*, **48**, 288, 1936.
[4] Wilson, F. N., F. D. Johnston, and P. S. Barker, *J. Clin. Investigation*, **16**, 664, 1937.

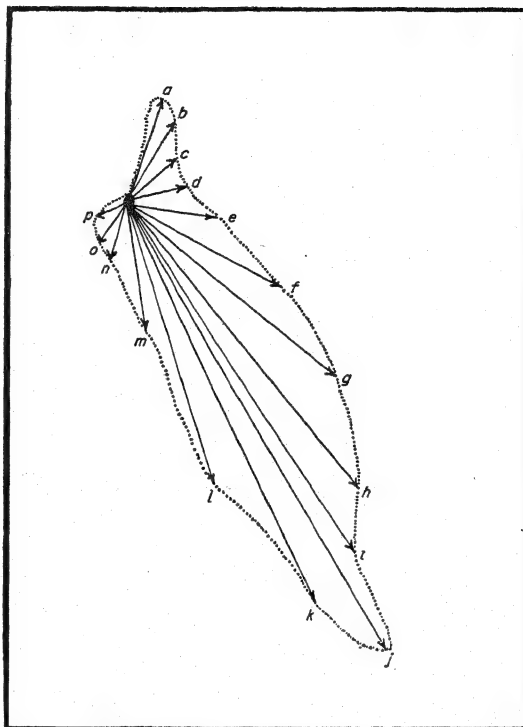



Fig. 57. Vectorcardiogram obtained, according to Mann, by joining the apexes of the vectors corresponding to the instantaneous electrical axes at different moments during the recording of QRS.

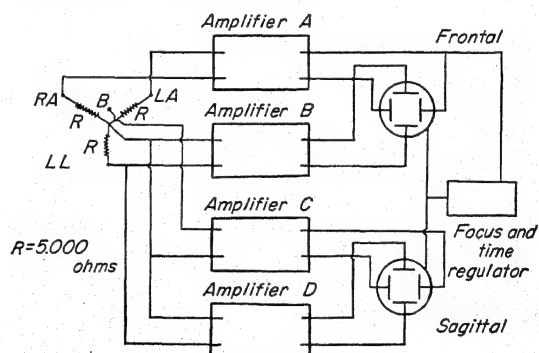



Fig. 58. Diagram of leads for spatial vectorcardiography with two cathode-ray oscillographs. *R*, resistance; *RA*, right arm; *LA*, left arm; *LL*, left leg; *B*, back.

Leads. The subject can be connected to the oscillograph in one of several ways, but in order

[1] Rochet, J., and M. Vastesaeger, *Trav. labor. Inst. Solvay*, **29**, 40 and 55, 1944; Sulzer, R., and P. W. Duchosal, *Helvet. physiol. pharmacol. acta.*, **4**, 285, 1946.

to obtain comparable results it would be necessary to adopt a uniform technique. This, however, has not yet been achieved, and each worker or group of workers uses the method he considers most advantageous. Everyone admits, however, that stereovectorcardiography gives more complete results; therefore the records should be obtained in at least two planes—simultaneously, with two oscillographs, or successively, if there is only one oscillograph. Usually vectorcardiograms in the frontal and sagittal planes are recorded, and sometimes a third one in the horizontal plane is also obtained.

Valid results are obtained in vectorcardiography with the cathode-ray oscillograph only when both electrodes are placed at the same distance from the heart and when this distance has a certain magnitude.[1] In man, the minimum distance should be 12 cm.[2] Several leads have been proposed for obtaining spatial vectorcardiograms. Some workers consider the heart to be embedded in an approximately regular tetrahedron,[3] or in a rectangular trihedron,[4] or in the center of a cube.[5] If the tetrahedron conception is adopted, three electrodes are placed one on each arm and the left leg as for the standard ECG, and a fourth on the back immediately to the left of the mid-line at the level of the seventh dorsal spinal process (Wilson *et al.*). A frontal-plane exploration is obtained with the electrodes on the right and left arm connected to the oscillograph plates which give horizontal deflections, and with one electrode on the left leg and the other a Wilson central terminal (see page 137) connected to the plates which give vertical deflections. A sagittal-plane exploration is made when a Wilson central terminal and the dorsal electrode are connected to the plates giving horizontal deflections, and a Wilson central terminal and the electrode on the left leg are connected to the plates giving vertical deflections. Figure 58 shows the connections when two oscillographs are used.

FIG. 59. Diagram of leads for spatial vectorcardiography according to the rectangular-trihedron hypothesis. (*Duchosal, P. W., and J. R. Grosgurin, Circulation,* 5, *237, 1952.*)

FIG. 60. Diagram of leads for vectorcardiogram in frontal plane. Electrodes *O* and *N* record the vertical component, and electrodes *O* and *P* the horizontal component. (*After Duchosal, P. W., and R. Sulzer, La Vectorcardiographie, S. Karger, Basel,* 1949.)

[1] GILLARD, G., V. HENDRICKX, and B. TACCARDI, *Acta cardiol.* (Brussels), **6**, 868, 1951.

[2] GRANT, R. P., *Circulation*, **1**, 878, 1950; LAMB, L. E., and E. G. DIMOND, *Am. Heart J.*, **44**, 165, 1952.

[3] WILSON, F. N., F. D. JOHNSTON, and C. E. KOSMANN, *Am. Heart J.*, **33**, 594, 1947; CONWAY, J. P., J. A. CRONVICH, and G. E. BURCH, *Am. Heart J.*, **38**, 537, 1949.

[4] DUCHOSAL, P. W., and J. R. GROSGURIN, *Circulation*, **5**, 237, 1952.

[5] GRISHMAN, L., E. R. BORUN, and H. L. HAFFE, *Am. Heart J.*, **41**, 483, 1951.

If the rectangular-trihedron conception is adopted (Duchosal and Grosgurin), four points are marked, three on the back and one on the chest, all at the same distance from the heart (Fig. 59). The frontal plane is explored with the electrodes *OP* connected with the plates giving horizontal deflections, and the electrodes *ON* connected with the plates giving vertical deflections (Fig. 60). A horizontal transverse plane at the level of the sternal margin of the fifth or sixth intercostal space is explored when the electrodes *OP* are connected with the plates giving horizontal deflections and the electrodes *OQ* are connected with the plates giving vertical deflections.

If the cube hypothesis (Grishman *et al.*) modified by Lamb and Dimond[1] is adopted, electrodes are placed on the right and left arm, a third one on the abdomen immediately above the umbilicus, another on the back at the level of the second lumbar vertebra, and lastly one on the right midaxillary line at the same level as the two preceding electrodes (Fig. 61). A frontal vectorcardiogram is obtained when the arm electrodes are connected with the plates giving horizontal deflections and the right-arm electrode and the midaxillary-line electrode with the plates giving vertical deflections. A sagittal vectorcardiogram is obtained when the right-arm electrode and the midaxillary electrode are connected to the plates giving vertical deflections and the abdominal and lumbar electrodes are connected with the plates giving horizontal deflections. A horizontal vectorcardiogram is obtained when the right- and left-arm electrodes are connected with the plates giving horizontal deflections and the umbilical and lumbar electrodes with the plates giving vertical deflections.

The polarity of the electrodes must be taken into account on making the connections. The deflections should be calibrated, and it should be possible to recognize the direction of displacement of the luminous spot on the screen.

FIG. 61. Electrode placement for spatial vectorcardiography, according to the cube hypothesis. (*After Lamb, L. E., and E. G. Dimond, Am. Heart J., vol. 44, p. 165, 1952.*)

Vectorcardiographic loops. A deflection in the oscillographic (or galvanometric) record due

[1] LAMB and DIMOND, *loc. cit.*

to electric currents arising in the heart is caused by a difference in the potentials at the points where the electrodes are placed. A dipole is, therefore, established as the resultant of all the electrical changes that occur. As there are several deflections of different amplitude and sign in the course of the cardiac cycle, the fundamental electrical phenomenon may be resolved into several dipoles of varying intensity and direction. Each one of the waves in the ECG is the expression of a dipole. The duration and shape show in each case that the vector alters in size, direction, and sign as time passes. A cathode-ray oscillograph connected to the body, as described above, can therefore show the evolution of several vectors, each one in the shape of a loop, the figure being always a closed one.

Usually vectorcardiograms show three principal loops corresponding to the development of the processes which in the ECG give the P, QRS, and T waves (Fig. 62). The symbols for these loops are P sÊ-loop, QRS sÊ-loop, and T sÊ-loop. The *axis of the loop* is the line joining its origin with its furthest point. It corresponds to the *instantaneous electrical axis* at that moment. Visualized in space, these loops should be considered as limiting plane areas.

The components P, QRS, and T of the vectorcardiogram can be represented in space by wire models of the corresponding loops which are mounted according to the spatial relations shown by the results obtained in the different planes. Graphic reproduction of these models is obtained by stereoscopic photographs.

FIG. 62. Normal vectorcardiograms. Right, frontal-plane projection; left, sagittal-plane projection in the same subject. P, QRS, and T loops are figured. Heavy line records slow movement; broken line, rapid movement. (*After Conway et al., Am. Heart J., vol. 38, p. 537, 1949.*)

Normal vectorcardiogram.[1] There are individual variations in the vectorcardiograms of normal subjects, but certain features are common to all (Fig. 62).

The axis of the P sÊ-loop is directed downward, slightly forward, and to the left. The anterior surface of the area it encircles is directed upward and to the left. It is recorded counterclockwise when facing the subject. Its contours vary more with respiration than those of QRS sÊ and T sÊ-loops; they lengthen, narrow, and become more vertical during inspiration.

The axis of the QRS sÊ-loop is directed downward, forward, and toward the left. The loop has an elliptical shape, its width being less than one-third its length. The anterior aspect of the area it encircles is directed upward and to the right. It is recorded clockwise when facing the subject, slowly at first, then faster throughout the major portion, and slowly again near its end.

The axis of the T sÊ-loop is directed downward, forward, and to the left. The loop has a very narrow elliptic shape. The anterior aspect of the surface it encircles is directed upward and to the right. It is recorded clockwise when facing the subject, slowly in the efferent portion, faster in the afferent portion.

[1] CONWAY, CRONVICH, and BURCH, *loc. cit.*

The ratio of the axes of P sÊ-loop, QRS sÊ-loop, and T sÊ-loop is approximately 1:10:2.

The vectorcardiogram can be traced from the ECG records in several leads, and reciprocally from the spatial vectorcardiogram it is possible to deduce the shape of the ECG components in the different leads, including the precordial leads[1] (Fig. 63).

FIG. 63. Vectorcardiogram and ECG. Above: frontal and transverse vectorcardiogram (accurate copies of original tracings). Numbers along vectorcardiogram, time in 0.01 sec. The T curve is drawn more darkly than the QRS; P is not represented. Scale of amplification is indicated on perpendicular axes of the vectorcardiograph figures (0.2 mv.). The broken lines represent the actual unipolar axes V_R, V_L, V_F, and V_1 to V_6. Dotted line connecting points 1 to 6 (V_1 to V_6) reproduces part of the transverse perimeter of the thorax at the level of the ventricular mass. Below: table of comparisons between real ECG (left in each of three columns) and ECG derived from the vectorcardiogram (right in each of three columns). (*Duchosal, P. W., and J. R. Grosgurin, Circulation, vol. 5, p. 237, 1952.*)

THE ELECTROCARDIOGRAPHIC LEADS

The recording of an ECG is the equivalent of exploring, by means of a recording galvanometer, the electric field created by the activity of the heart. An electric field is established when an electric potential arises in a conducting medium. It is defined by lines of flow and equipotential planes. The lines of flow show the passage of current, *i.e.*, the paths along which electrons travel owing to the existence of different potential levels. Equipotential planes are perpendicular to the lines of flow and correspond to the different points in the field which are at the same potential level.

In the case of a dipole in a homogeneous weak conducting medium, points at the same potential form spheres. A section through a field arising in pointlike positive and negative poles placed near to each other shows isoelectric

[1] DUCHOSAL and GROSGURIN, *loc. cit.*

lines (sections of the corresponding spheres) distributed as in the diagram in Fig. 64. The lines of flow are not represented. The line *I-I* corresponds to points at zero potential due to reciprocal neutralization of charges of opposite sign. All points on the right of this line will have a certain charge; points on the left will have charges of the opposite sign. The potential (both positive and negative) will decrease as the point considered lies farther from the source until it falls to zero for points at infinity.

Electromotive forces arising during cardiac activity can be resolved at each moment into a dipole with an electric field which does not differ much from that represented in Fig. 64.The main difference would be a less regular distribution of the isoelectric lines due to differences in conductivity of the different tissues surrounding the heart and to the limitation of the conducting medium by the body surface.

If two points in the field are connected by a conductor (a metal of much lower resistance than that of the medium in which the dipole is situated), a current will flow along it if the points are at different potential levels. A galvanometer inserted in the conductor will show the passage of current and its variations, thus permitting the exploration of the electric field. A recording galvanometer gives records of current deflections which permit a more detailed analysis. By placing the ends of the conductor (electrodes) at different points, a clear idea of the electric field can be obtained.

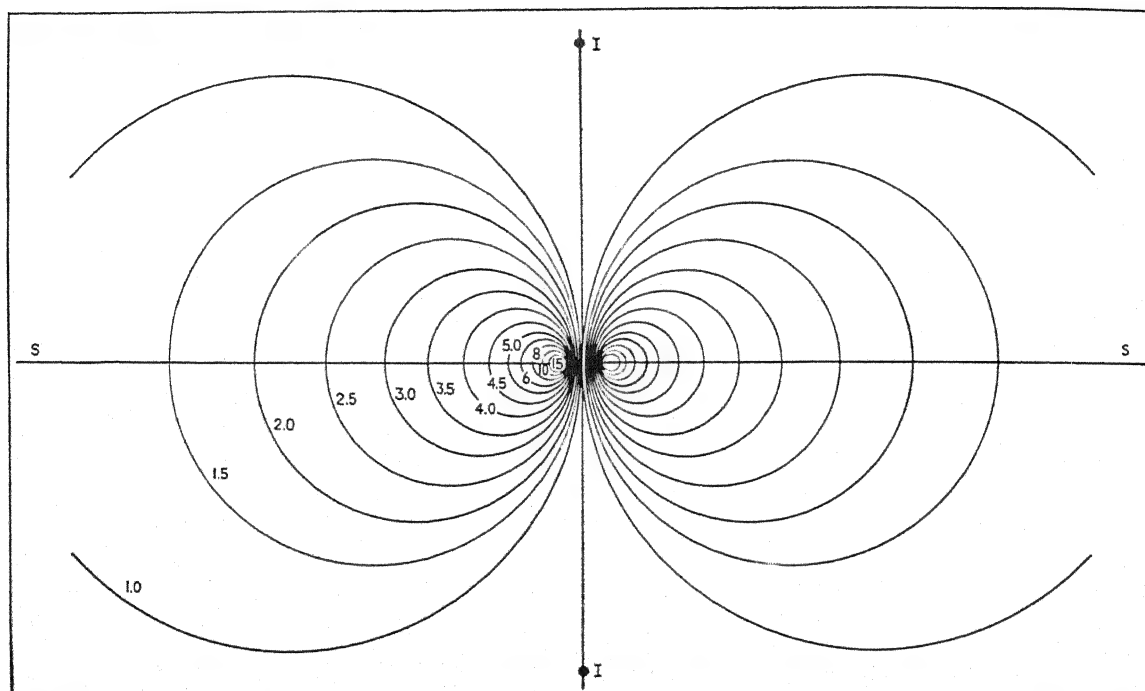



FIG. 64. Electric field of a dipole. Distribution of isoelectric planes created by a dipole in a homogeneous unlimited tridimensional medium. S, S, axis of the dipole; I, I, plane of zero potential. Figures to the left of I, I, values for positive charges; corresponding circles to the right of I, I, have the same absolute values but of opposite sign.

There is no object in examining all the possibilities, but the current will be at a maximum when the electrodes are placed on the axis *S-S* (Fig. 64) of the dipole, since the difference in potential is highest for points on this line; there will be no passage of current if the electrodes are placed on the line *I-I* perpendicular to the axis *S-S* and passing through the center of the dipole. If the electrodes are placed on lines parallel to the axis of the dipole, a current will pass which will be greater (for a given distance between the electrodes) the nearer the line is to the axis, and will be at a maximum when the electrodes are placed directly on the poles (maximum positivity and negativity). On the contrary, there will be no passage of current if the electrodes are placed on lines perpendicular

to the axis of the dipole at points situated at the same distance from the poles because these lines join isoelectric points.

The electric field produced by electric potentials arising in the course of cardiac activity is not stationary, as was explained when discussing the electrical axis of the heart. The dipole that arises is continually changing in potential and in the direction of its axis (the line joining the poles). At times it is completely reversed; thus the dipoles corresponding to the deflections Q or S are orientated in the direction opposite to that giving rise to the R wave.

Direct leads. When the electrodes are placed directly on the surface of the heart, the derivation is called a direct lead. This is a valuable method in experimental work, but can be applied to human beings only in exceptional cases, *e.g.*, in the course of intrathoracic operations.

Indirect leads. The cardiac electric field is usually explored by placing the electrodes on the body surface (indirect leads).

Standard limb leads. Einthoven recommended certain leads which since then have come into universal usage. The electrodes are placed on the limbs, and usually three combinations are used: lead I, right arm–left arm; lead II, right arm–left leg; and lead III, left arm–left leg. These three leads give a general idea of the cardiac potentials. They have already been referred to when describing the deflections in the ECG and when the electrical axis of the heart was discussed.

Precordial leads. A more localized analysis of cardiac potentials can be made by placing one electrode on the chest in the precordial area, and the other in some distant part of the body. This way of exploring the heart by placing one electrode near the source of the potentials is similar to the semidirect exploration made by placing one electrode directly on the surface of the heart (exploring electrode). The near electrode can thus capture potentials which owing to their small magnitude cannot be registered by the distal electrode.

An indefinite number of chest leads might be used, but in order to simplify the technique and to obtain comparable results, the exploring electrode is usually placed on one of the six following points:

1. On the right sternal margin at the fourth intercostal space.
2. On the left sternal margin at the fourth intercostal space.
3. Midway between the left sternal margin and left mid-clavicular line, at the level of a line going from the anterior end of the fourth intercostal space and the apex beat.
4. On the mid-clavicular line at the level of the apex beat.
5. On the left anterior axillary line at the level of the apex beat.
6. On the left midaxillary line at the level of the apex beat.

If the distal electrode is placed on the right arm, the precordial lead is called CR; it is CL if it is placed on the left arm, and CF if it is placed on the left leg. The precordial point explored is noted by a subscript using the figures given above, *e.g.*, CR_4, CF_5, etc. The galvanometer should be connected so that a positive charge in the precordial region provokes an upward deflection in the record. This result is obtained by connecting the precordial lead to the cable used for the left arm in the instruments commonly in use.

Wilson's central terminal. Wilson has suggested the use of a common terminal connected through 5,000-ohm resistances to each one of the right-arm, left-arm, and left-leg electrodes. The leads in this case are called CV with a subscript for the precordial point explored, *e.g.*, CV_2.

The object of Wilson's central terminal is to counterbalance potentials arising in the limbs so that the potential in the conductor remains sufficiently constant to be considered as zero during any moment of the cardiac cycle. If this were so, electrocardiograms obtained by this method would register only potential changes produced under the exploring electrode and would be similar to those given by a semidirect lead. Recent studies, however, have shown that derived potentials from any part situated outside the cardiac mass, even when very close to the heart, are due to the total resultant cardiac dipole[1] and that precordial leads reflect the mean direction in space of all the cardiac potentials, without any significant dominance on the sense of the deflection by that part of the myocardium which is immediately under the electrode.[2] It has been maintained that a unipolar lead is never obtained in electrocardiography.[3]

[1] Duchosal, P. W., and L. Grosgurin, *Acta cardiol.*, **4**. 425, 1949.
[2] Grant, *loc. cit.; ibid.*, **2**, 676, 1950.
[3] Spang, K., *Ztschr. f. Kreislauff.*, **38**, 405, 1949.

Limb unipolar leads. Following the ideas of Wilson and his associates, the study of so-called "limb potentials" has been much in vogue. One of the poles of the galvanometer is connected successively to each one of the limbs (right arm, left arm, left leg) so that a positive potential gives an upward deflection, and the other pole is connected to a Wilson central terminal, and the corresponding records are obtained. Goldberger disconnects from the common terminal the limb that is being explored and does not employ the usual resistances, thus simplifying the technique without altering the results. If Wilson's method is used, the leads are called VR, VL, and VF for the right arm, left arm, and left leg respectively. If Goldberger's technique is used, the letter a (meaning augmented, because the resistances have been removed from the common terminal) precedes the foregoing nomenclature, *i.e.*, aVR, aVL, aVF.

Intracardiac leads.[1] Exploration of the cavities of the heart by the introduction of a catheter through a vein or an artery has often been performed in man. Sometimes it is complemented by recording the intracardiac electrogram.

The catheter in the auricle or the ventricle is connected with the exploratory electrode, and the circuit is completed by means of a Wilson central terminal.

Einthoven's law. Einthoven deduced from a mathematical analysis of his triangle that the height of R in lead II should be equivalent to the algebraic summation of the heights of R in leads I and III. Usually the ECG is obtained by successive, not simultaneous, recording in the three leads; for this reason the experimental results are in only approximate, but nevertheless fairly satisfactory, agreement with the theoretical predictions.

Interpretation of records obtained with precordial leads. According to Wilson and his associates[1] precordial leads serve much the same purpose in the clinical exploration of the heart as the direct unipolar leads introduced by Lewis and Rothschild[2] in the experimental study of the spread of excitation in the dog's heart. The process of excitation by means of these leads can be examined at the point at which it arises, and as it spreads through the heart until it is completed. The method is especially valuable for the exact diagnosis of ventricular hypertrophy, bundle-branch block, and coronary occlusion. Records taken from different points must be compared in order to obtain useful information; the record given by a single precordial lead has little or no significance. The number of precordial leads used should be such as to give adequate data about the electrical changes taking place in both ventricles. All the six leads mentioned should be used, and in certain cases others should be added.

Records obtained with precordial leads can be interpreted by comparing them with those obtained by unipolar exploration of the surface of the heart in experimental animals. The spike of the main upward deflection of the rapid ventricular complex (R), according to Wilson and his associates, marks the beginning of the intrinsic electrical potential, *i.e.*, that occurring in the spot under the electrode; therefore it is a sign of excitation of the cardiac fibers in this place. If this occurs early in the cycle the R wave is usually small and is followed by a deep S. R becomes larger and S smaller as excitation takes place later, and sometimes a Q wave precedes R. The aspect of the QRS complex is, therefore, determined by the way in which the fibers under the electrode are excited, in relation to the spread of excitation throughout the whole myocardium.

Precordial lead records in normal subjects. Records obtained with electrodes placed on the right side of the precordial area show a small R wave, which soon reaches its peak, followed by a deep S (Fig. 65, CV_1). According to Wilson this type of record is due to the fact that the excitatory state, spreading outward, reaches the surface of the right ventricle before it does that

[1] Lenegre, J., and P. Maurice, *Paris méd.*, **35**, 23, 1945; *Arch. d. mal. du coeur*, **38**, 298, 1945; Hecht, H., *Am. Heart J.*, **32**, 39, 1946; Battro, A., and H. Bidoggia, *Am. Heart J.*, **33**, 604, 1947; Sodi-Pallares, D., M. Vizcaine, J. Soberon, and E. Cabrera, *Am. Heart J.*, **33**, 819, 1947; Kossman, C. E., A. R. Berger, J. Brumlik, and S. A. Briller, *Am. Heart J.*, **35**, 309, 1948; Schlesinger, P., A. B. Benchimol, and M. R. Cotrim, *Arq. d. clin.*, **6**, 3, 1948; Duchosal, P. W., C. Ferrero, J. P. Dovet, P. Adereggen, and B. Rilliet, *Cardiologia*, **13**, 113, 1948; Levine, H. D., *et al.*, *Am. Heart J.*, **37**, 46 and 64, 1949; Kossman, C. E., *et al.*, *Circulation*, **2**, 10, 1950. Zimmerman, H. A., and H. K. Hellerstein, *Circulation*, **3**, 95, 1951.

[1] Johnston, F. D., E. F. Rosenbaum, and F. N. Wilson, *Mod. Concepts of Cardiovas. Dis.* (American Heart Association), **12**, Nos. 6 and 7, 1943; Wilson, F. N., *et al.*, *Am. Heart J.*, **27**, 19, 1944.

[2] Lewis, T., and M. A. Rothschild, *Phil. Trans. Roy. Soc.*, London, **B206**, 181, 1915.

of the left ventricle because of the greater thickness of the wall of the latter. Leads taken from the left side show a large R, with a late peak, frequently preceded by a small Q, and followed by a small S (Fig. 65, CV_6). The Q wave (when it exists) is apparently due to registration of the potential in the cavity of the ventricle (which remains negative throughout the QRS complex) at a moment in which the fibers between the cavity and the electrode have not yet entered into activity. Ventricular complexes of records taken with the electrode placed in the center of the precordial area are intermediate between the two former types (Fig. 65, CV_3). The situation and dimensions of this intermediate area vary in different subjects; its center is usually under lead C_3. The causes determining these variations are still unknown.

FIG. 65. Electrocardiograms obtained from a normal subject with the usual (upper three records) and precordial leads. Wilson's terminal was used. Note how R increases and S diminishes as the exploring electrode is displaced from right to left (CV_1 to CV_6) on the precordial area. (*Courtesy of Dr. J. L. Gilbert.*)

The T wave in the precordial leads of normal adults may be inverted in C_1, but not in records with the other leads.

The duration of the QRS complex is always less than 0.1 sec. in normal subjects. The peak of R in records of leads on the right usually precedes the peak of R in left-side leads by 0.02 sec.

Precordial lead records in abnormal conditions. In certain pathologic conditions records taken with precordial leads show variations with respect to those of normal subjects.

Hypertrophy of the left ventricle causes the following changes to appear:

1. The maximum voltage of the largest QRS deflection is, on an average, greater than in normal subjects.
2. QRS lasts 0.10 to 0.11 sec.
3. Leads from the right side of the precordial area give a small R wave, which may even be absent.
4. The intermediate area is shifted to the left.
5. Leads from the left side give a large R, with late peak, and the T wave is usually inverted.

Hypertrophy of the right ventricle causes changes in the opposite direction of those just mentioned. Right-side leads give a very large R wave, some-

times preceded by a Q wave; S is small or absent, and T is usually inverted. Records from the left side are similar to those given by leads placed on the right side in normal subjects, *i.e.*, R is small or absent, S is large, Q is missing, and T is an upward deflection.

Intraventricular block gives records similar to those of ventricular hypertrophy, but in bundle-branch block the QRS complex always lasts more than 0.11 sec. and the main wave (R or S) has a wide base and is split or splintered.

In cases of *left bundle-branch block*, right precordial leads show a large descending wave, frequently splintered, sometimes preceded by a small R with an early peak. The small R and large S are apparently due to early stimulation of the right ventricle, and the long duration of the QRS complex is due to the delay in stimulation of the left ventricle. Left-side lead records (CV_5 and CV_6) show a single split or splintered R wave with a wide base. The intermediate area is shifted to the left. The T wave is usually deflected in the opposite direction from that of the main wave of the QRS complex, *i.e.*, upward (positive) in right-side leads and downward (negative) in left-side leads.

In cases of *right bundle-branch block*, right-side leads show an R wave varying in height (the descending limb of which may not reach the base line), followed by a second wave, known as R′. The aspect is that of an M with the second peak higher than the first. As the electrode is displaced toward the left, the height of the first wave increases and R′ becomes smaller. Left-side leads show a single high and narrow R wave followed by a deep S, and sometimes preceded by a Q wave. According to Wilson, the R wave in right-side leads is caused by excitation of the ventricular septum, and R′ to delayed excitation of the right ventricular wall under the electrode at a moment when electric forces of opposite sign are not arising in other parts of the myocardium.

Coronary occlusion does not always produce changes in the ECG taken with the usual leads even when it causes an infarct in the myocardium; therefore precordial leads are particularly useful in the diagnosis of this condition. Typical changes are

1. Shifting of the ST segment above or below the isoelectric base line.
2. Negative T wave.
3. Deep Q waves with a wide base (0.04 sec. or more).

The precordial ECG varies in the course of the disease in the same way as the usual ECG. Deviation of the ST segment occurs immediately, and disappears after a few days or weeks; changes in the T wave may persist for months.

Table 17. Electrocardiographic Position of the Heart

Position	*QRS in*	*Shape of QRS*
Vertical.........	VL	Similar to those in CV_1 and CV_2
	VF	Similar to those in CV_5 and CV_6
Semivertical.....	VL	Small
	VF	Similar to those in CV_5 and CV_6
Intermediate.....	VL and VF	Similar to those in CV_5 and CV_6
Semihorizontal...	VL	Similar to those in CV_5 and CV_6
	VF	Small
Horizontal.......	VL	Similar to those in CV_5 and CV_6
	VF	Similar to those in CV_1 and CV_2

According to Wilson and associates.

Interpretation of records obtained with unipolar limb leads. *Electrocardiographic position of the heart.* The increasingly frequent use of monopolar leads has given rise to the idea that deflections registered by leading off cardiac potentials from a point on the surface of the body are similar to those which would be obtained if the electrode were placed directly on the surface of the heart at the spot nearest the exploring electrode. This idea and the possibility of using an indifferent electrode such as Wilson's common terminal have led to the belief that records of potentials in each one of the limbs (right arm, left arm, left leg) during the cardiac cycle might give important data on the orientation of the heart.

In the normal position, potentials arising in the ventricular cavities, which are negative, are regularly transmitted to the right arm, and deflections arising in the right and left ventricular surface are transmitted in varying degrees to

the left leg and left arm. The shape of the potentials registered in the left arm and leg ultimately will depend on the position of the heart with respect to its own longitudinal axis. As the heart becomes more horizontal, transmission of potentials from the surface of the right ventricle to the leg increases. According to Wilson and his associates, monopolar leads from the limbs give records which define five electrocardiographic positions of the heart (Table 17).

Unipolar limb leads and standard limb leads. The three standard leads I, II, and III are closely related to the three monopolar leads from the limbs. If the values of any two are known, the values of the other four can be calculated.

The sum of the potentials from the three limbs, represented by the angles of Einthoven's triangle, equals zero at any moment of the cardiac cycle, *i.e.*, VR + VL + VF = 0. Therefore, taking into account the conventional polarity used in obtaining the records,

$$L_{I} = VL - VR$$
$$L_{II} = VF - VR$$
$$L_{III} = VF - VL$$

Other applications of monopolar limb leads. Monopolar limb leads can give valuable information for the diagnosis and localization of myocardial infarcts. In some cases the characteristic changes appear in only one of these monopolar limb leads.

PRACTICAL SIGNIFICANCE OF THE ELECTROCARDIOGRAM

A careful study of the electrocardiogram gives information on the spread of excitation to the different chambers of the heart. It tells where the impulse originated and whether it spread along normal or abnormal paths, at normal or abnormal speed. It gives no information whatsoever on the mechanical vigor of the heartbeat. The electrocardiogram is of real value in the establishment of a correct diagnosis in cases of abnormal cardiac rhythm and in cases of myocardial damage resulting from impaired blood supply or from toxic or infectious conditions.

BIBLIOGRAPHY

Ashman, R., and E. Hull, "Essentials of Electrocardiography," 3d ed., Macmillan, New York, 1945.

Barker, J. M., "The Unipolar Electrocardiogram," Appleton-Century-Crofts, New York, 1952.

Burch, G., and T. Winsor, "A Primer of Electrocardiography," Lea & Febiger, Philadelphia, 1945.

Cabrera, E., "Bases électrophysiologiques de l'électrocardiographie," Masson et Cie, Paris, 1948.

Craib, W. H., *Heart*, **14**, 71, 1927 (the electric field surrounding the active heart).

Doumer, R., "Les principes de l'électrocardiographie," Masson et Cie, Paris, 1950.

Duchosal, P. W., and R. Sulzer, "La Vectocardiographie," S. Karger, Basel and New York, 1949.

Einthoven, W., in Bethe's "Handbuch der normalen und pathologischen Physiologie," **8–2**, 785, Springer, Berlin, 1928.

Frédéricq, H., "Aspects actuels de la physiologie du myocarde," Presses Universitaires de France, Paris, 1927.

Goldberger, E., "Unipolar Lead Electrocardiography," 2d ed., Lea & Febiger, Philadelphia, 1949.

Grant, R. P., and E. H. Estes, Jr., "Spatial Vector Electrocardiography. Clinical Electrocardiographic Interpretation," Blakiston, Philadelphia, 1951.

Graybiel, A., "Clinical Electrocardiography," Nelson, New York, 1951.

Grishman, A., and L. Scherlis, "Spatial Vectorcardiography," Saunders, Philadelphia, 1952.

Groedel, F. M., and P. R. Borchardt, "Direct Electrocardiography of the Human Heart and Intrathoracic Electrocardiography," Brooklyn Medical Press, New York, 1948.

Hecht, H. H., "Basic Principles of Clinical Electrocardiography," Charles C Thomas, Springfield, Ill., 1951.

Jouve, A., "La Vectocardiographie en clinique," Masson et Cie, Paris, 1950.

Katz, L. N., "Electrocardiography," 2d ed., Lea & Febiger, Philadelphia, 1946.

Katz, L. N., Clinical Electrocardiography. Its Present Position and Possible Potentialities, *Circulation*, **2**, 94, 1950.

Lepeschkin, E., "Modern Electrocardiography," Williams & Wilkins, Baltimore, 1951.

Lewis, T., "Mechanism and Graphic Registration of the Heart Beat," 3d ed., Shaw and Sons, London, 1925.

Lewis, T., "Clinical Electrocardiography," 4th ed., London, 1928.

Pardee, H. E. B., "Clinical Aspects of the Electrocardiogram," 4th ed., Hoeber, New York, 1941.

Rosenbaum, F. F., "Clinical Electrocardiography," Oxford, New York, 1950.

Schaefer, H., "Das Electrokardiogram, Theorie und Klinik," Springer, Berlin, 1951.

Schaffer, A. I., "Cardiography in General Practice. Electrocardiography, Vectorcardiography, and Ballistocardiography," Williams & Wilkins, Baltimore, 1952.

Scherf, D., and L. J. Boyd, "Clinical Electrocardiography," Mosby, 3d ed., Saint Louis, 1949.

SODI-PALLARES, D., "Nuevas Bases de la Electrocardiografía," 2d ed., La Prensa Médica Mexicana, México, D.F., 1951.

STROUD, W. D., "The Diagnosis and Treatment of Cardiovascular Disease," Davis, Philadelphia, 1950.

WEBER, A., "Die Electrokardiographie und andere graphische Methoden in der Kreislaufdiagnostik," 2d ed., Springer, Heidelberg, 1947

WIGGERS, C. J., "Principles and Practice of Electrocardiography," Mosby, Saint Louis, 1929.

WILSON, F. N., "The Precordial Electrocardiogram," Recording and Statistical Corp., New York, 1943.

WILSON, F. N., A. G. MACLEOD and P. S. BARKER, "The Distribution of the Currents of Action," etc. University of Michigan Studies, Scientific Series 10, Ann Arbor, 1933

CHAPTER 16

The Excitation of the Heart

RHYTHMICITY IS ONE of the fundamental and more remarkable properties of the heart. It is characterized by the capacity of the heart to generate within itself periodic impulses that will determine its activity.

Rhythmicity is inherent in the cardiac muscular fiber, but it is not equally developed either in all fibers or in the different parts of the heart. If we consider this organ as a whole, it is obvious that the portion with more developed rhythmicity—the one, therefore, that will be able to generate more impulses in unit time—will impose its own rhythm on the entire organ. This is what actually happens.

NORMAL EXCITATION OF THE HEART

The sinoauricular (S-A) node. The normal pacemaker. In mammals, under normal conditions, the impulses that determine the heartbeat are originated in the sinoauricular (S-A) node, a vestigial remnant of the sinus venosus of fishes and amphibians. It is formed by a relatively small mass of cardiac muscle showing a peculiar histologic aspect and with certain chemical characteristics. This structure is called the "normal pacemaker" of the heart because of its particular significance in cardiac activity; it is located in the sulcus terminalis, formed by the superior vena cava and the posterior wall of the right auricle. It is also known as "Keith and Flack's node," after its discoverers.[1]

The fibers of the S-A node are less differentiated than the ordinary fibers of the myocardium, the striation is less marked, and the glycogen content is lower. They are surrounded by dense connective tissue.

It is relatively easy to demonstrate that the S-A node is the normal pacemaker of the heart. This can be done in several ways, all equally convincing.

Changes in temperature localized on the S-A node modify the heart rate; they do not produce this effect when they occur in any other part of the heart. Thus if a test tube filled with warm or cold water is placed on different parts of the heart, there is no change in heart rate except when the tube is placed on the sulcus terminalis, *i.e.*, near the S-A node. In this case, if the tube contains warm water, the heart rate will increase, and it will diminish if the water is cold.

Exploration of the heart with adequate electrodes shows that the first place in which potential differences occur is the region of the S-A node.

If the S-A node is extirpated, the cardiac rhythm changes; the heart rate diminishes, and auricles and ventricles seem to beat simultaneously. Electrocardiograms recorded under the new circumstances will also show that the heart is beating with an abnormal rhythm. This experiment proves, furthermore, that the S-A node is not the only center of cardiac rhythmicity.

The S-A node is therefore the part of the heart which, under normal conditions, because of its peculiar metabolism, originates the greatest number of impulses in unit time, thereby imposing its own rhythm on the rest of the heart. The normal heart rhythm is known as the sinus rhythm, because of its origin. Its basal frequency of discharge is 75 to 80 impulses per minute.

Auricular excitation. The impulse originated

[1] KEITH, A., and M. FLACK, *J. Anat. & Physiol.*, **41**, 172, 1907.

in the S-A node spreads to the auricle and is propagated in all directions, so that all the fibers are excited. Special paths for the conduction of the impulse have not been anatomically differentiated. Nevertheless, electrical exploration has shown that the impulse travels more rapidly

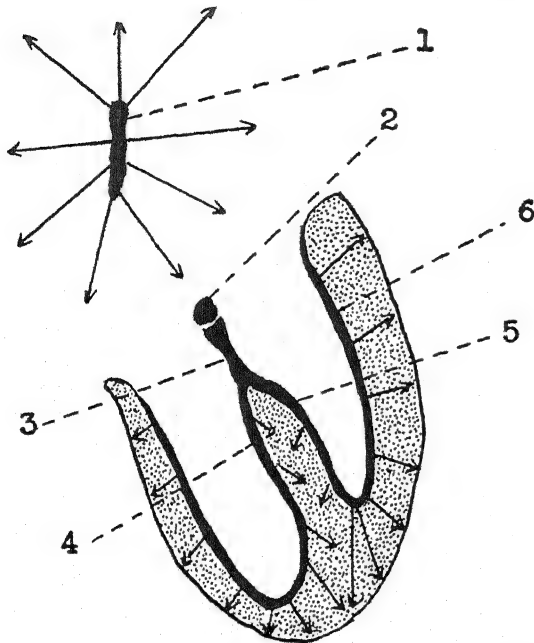



Fig. 66. Diagram of the excitation and conduction system of the heart. 1, sinoauricular node; 2, auriculoventricular node; 3, bundle of His; 4, right branch of the bundle of His; 5, left branch of the bundle of His; 6, Purkinje's system. The arrows indicate the direction in which excitation spreads through the auricular and ventricular muscle.

toward the left auricular appendage and toward the A-V node than to other parts of the auricle. In the dog, Lewis observed velocities of 0.5 to 1.2 m. per sec. in the different directions. In man direct exploration of the heart, using Lewis's method, has shown that excitation spreads at 1.8 to 2.7 m./sec., *i.e.*, twice the speed found in dogs.[1]

The auriculoventricular (A-V) node. In the posteroinferior part of the interauricular septum, near the opening of the coronary sinus, there is a muscular formation known as the auriculoventricular node or the Aschoff-Tawara node.[2] It is an important center of rhythmicity

[1] Kossman, C. E., J. Brumlik, and S. Briller, *J. Clin. Investigation*, **26**, 1186, 1947.

[2] Tawara, S., "Das Reizleitungssystem des Saugetierherzens," Jena, 1908.

and the origin of the bundle of His.[1] This muscular bundle forms a connecting bridge between the auricles and the ventricles and is therefore of special physiologic importance.

The A-V node is made up of two parts, anatomically separated and embryologically different. The upper, auricular, or coronary part develops from the primitive sinus tissue, and its fibers, like those of the S-A node which also have the same origin, have a low glycogen content. The lower, or ventricular part of the node evolves from the auricular canal; its fibers are larger and richer in glycogen than those of the upper part. The bundle of His emerges without any transition from the ventricular portion of the A-V node.

In normal conditions the A-V node acts as a relay station for the impulses originated in the S-A node. On reaching the A-V node the impulses are delayed as if they had to overcome an obstacle. The speed of the impulse in this region has been calculated to be 0.2 m. per sec. (in the dog). This delay permits auricular contraction to end before excitation of the ventricles begins.

In abnormal conditions, when the normal pacemaker is not functioning, or when the rhythmicity of the A-V node is increased to a level above that of the S-A node, the A-V node takes over the functions of pacemaker. The prevailing rhythm is then called an A-V nodal rhythm. When it is established because the S-A is not functioning, the heart rate is about 50 beats per minute, and auricles and ventricles beat almost simultaneously. The nodal rhythm will be considered in greater detail further on.

Ventricular excitation. From the A-V node the impulse spreads rapidly through a specialized connecting system made up of functionally differentiated myocardial fibers, thus reaching all the muscle fibers of the ventricles. The conducting system is formed by the bundle of His, its branches, and the Purkinje network.

The bundle of His is continuous with the ventricular portion of the A-V node. It passes through the A-V septum and goes forward and downward to the membranous portion of the interventricular septum. It soon divides into two branches, right and left, one for each ventricle. These branches subdivide many times and thus make multiple connections with a subendocardiac network formed by special fibers,

[1] His, W., *Arb. a. d. med. Klinik*, **14**, Leipzig, 1893.

histologically and chemically similar to those constituting the bundle of His. They are called Purkinje fibers. Offshoots from this network penetrate between the ordinary muscle fibers of the ventricular walls. Figure 66 is a diagrammatic representation of the conducting system in the heart of mammals. The A-V node and the bundle of His and its branches have been demonstrated by dissection in the human heart.[1]

The impulse spreads through the bundle of His, its branches, and the Purkinje network at a speed of 5 m. per sec. (in the dog). This great speed results in a very rapid entrance into activity of all the ventricular muscle fibers. The whole process takes about 0.07 sec. in the dog (Lewis). The speed of conduction in the contractile myocardium is around 0.5 m. per sec., *i.e.*, ten times less.

The duality of the myocardium. In the heart there is therefore a complete system made up of nodes and special pathways, which originates and transmits the impulse successively and in an orderly way to the different parts of the organ. The fibers of this system, as already stated, have peculiar histologic and chemical features. There are therefore two types of myocardium: one constituting the excitation and conduction system, with the properties of rhythmicity and conductivity especially developed in its fibers, and another (the common myocardium) making up most of the muscular walls of the heart, with contractility as its most prominent property (contractile myocardium).

Varieties of sinoauricular rhythm. In man, under normal conditions, the S-A pacemaker originates and sends out 75 to 80 impulses per minute at remarkably constant intervals, thus determining the normal heart rate. This can be modified by the isolated or joint action of several factors, such as nerve impulses, hormones, drugs, temperature changes, etc. The number of impulses originated in unit time can either increase, in which case there will be a sinus tachycardia, or decrease, in which case there will be a sinus bradycardia.

The impulses are not always sent out from the S-A node with perfect regularity. During inspiration very often the heart rate is greater than during expiration. This sinus arrhythmia, or respiratory arrhythmia, is not abnormal; it is due to nervous influences carried to the heart by its extrinsic nerves, whose nuclei of origin are subject to the variable activity of the respiratory center. This type of arrhythmia is frequent in children and young people, and a marked sinus arrhythmia is almost constantly observed in the dog.

[1] WIDRAN J., and M. LEV, *Circulation*, **4**, 863, 1951.

In all cases of sinus arrhythmia, the heartbeat is normal in so far as the order in which excitation spreads to the different parts of the heart is concerned; neither is the nature of the contraction altered, only the duration of the cycle—specifically, ventricular diastole.

The electrocardiogram always shows a complete set of normal deflections (P, Q, R, S, and T) at each cycle, without any change in shape, in sequence, or in the intervals between them. The only changes consist in a greater frequency in tachycardia, or a lesser one in bradycardia, and the irregularity of the intervals between the complexes, should a sinus (respiratory) arrhythmia be present.

The changes in the sinus rhythm are not necessarily abnormalities. Changes in heart rate are usually physiologic mechanisms of adaptation to the conditions in which the heart is working (exercise) or the result of emotional states.

ABNORMAL EXCITATION OF THE HEART

The normal excitation of the heart is due to impulses originated in the S-A node and transmitted by the pathways that have just been described. This is also the mechanism that assures the greatest efficiency of the heartbeat. There are, however, many causes that frequently disturb the process of excitation in different ways, thus determining more or less important alterations in cardiac and circulatory dynamics. The fundamental aspects of abnormal rhythms will be considered here, because of the great contributions experimental physiology has made in throwing light upon their mechanism, and also because the normal functions of the heart will be more clearly understood through the analysis of the many resources and mechanisms available within itself to assure its efficiency, even when the normal processes have been greatly disturbed.

Abnormal Heart Rhythms

The sinus rhythm is sometimes called the normotopic rhythm. Rhythms originated in

other parts of the heart are called ectopic rhythms. If the abnormal rhythm is originated in some part of the excitation-conduction system it is called homotopic, and heterotopic when it originates outside this system (Hering). Ectopic rhythms develop when the excitability of the normal pacemaker is either depressed or abolished, or when excitability is abnormally increased at some lower focus. Pathological studies suggest, as an important cause of these disturbances, nutritional alterations due to vascular lesions affecting the involved region.

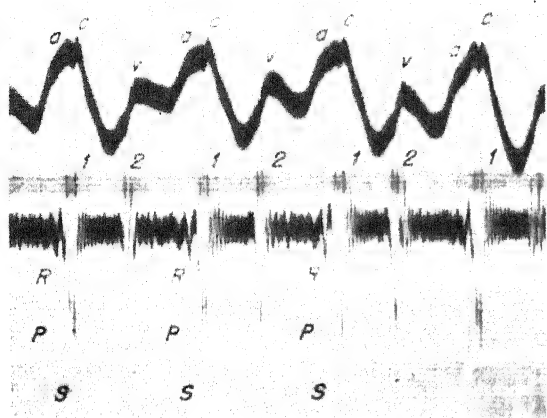



FIG. 67. Upper nodal rhythm (records of a human case). Phlebogram, phonocardiogram, electrocardiogram, and time in 0.2 sec. Note the brief duration of the PR interval in the ECG, the closeness of *a* to *c* in the venous pulse, and the reinforcement of the initial vibrations of the first heart sound. The auricular sound and the normal third heart sound are also registered. In the two last beats the second sound is reduplicated (split).

A-V nodal rhythm. If the S-A node is depressed or has lost its activity, the A-V node takes over the command of cardiac activity and an A-V nodal rhythm is established. The A-V node being a complex one, several types of A-V nodal rhythm may develop.

A nodal rhythm may develop in the upper, middle, or lower part of the A-V node. Such rhythms are called upper, middle, or lower nodal rhythms respectively. They are easily recognized in the ECG (Figs. 67, 68, 69). In cases of upper nodal rhythm the P wave (sometimes inverted) immediately after its commencement is hidden by the QRST ventricular complex. The interval between the beginning of P and the beginning of the ventricular complex (PR interval) is reduced, varying from 0.02 to 0.04 sec. (Fig. 67). In cases of lower nodal rhythm P (usually inverted) appears between the initial ventricular complex (QRS) and the final wave T, the P wave being nearer to the initial complex than to the final one. Finally in cases of middle nodal rhythm P is marked at the same time as QRS and is therefore "absorbed" and concealed by the initial ventricular complex; there is no P wave in the electrocardiogram (Figs. 68 and 69). Nodal rhythms are usually very regular.

If the A-V nodal rhythm results from the absence or functional depression of the S-A pacemaker, the heart rate is low, around 40 to 50 beats per minute. If, on the contrary, it is due to an increased rhythmicity of the A-V node, the frequency of discharge of the latter is greatly enhanced and an A-V nodal tachycardia is the result.

The A-V node can send out occasional impulses when the preponderant rhythm is of S-A origin. A-V nodal premature contractions (extrasystoles) are thus brought about; they have all the mechanical and electrocardiographical features of the A-V nodal beats. These extrasystoles can be originated in the upper, lower, or middle portions of the A-V node.

Idioventricular rhythms. When the normal S-A impulses, or any other stimulus of supraventricular origin, cannot spread to the ventricle because of a functional disturbance or on account of anatomical discontinuity of the His bundle, the ventricles do not stop beating. A center of rhythmicity may arise in any part of the ventricular myocardium (most frequently in the His bundle below the site of interruption) and become the pacemaker for the ventricles. The rhythm thus established is called an idioventricular rhythm. When beating with this type of rhythm the heart rate is always slow, around 30 beats per minute. Idioventricular rhythms will be considered in greater detail when dealing with heart block, a condition with which they are always associated.

Disturbances in Impulse Conduction

When the capacity to conduct impulses is diminished or suppressed in any part of the conducting system, it is said to be blocked. The degree and localization of block varies and can cause a greater or lesser disturbance in the activation of the heart. The principal types of

block are the following: (*a*) sinoauricular (S-A) block; (*b*) auriculoventricular (A-V) block; (*c*) bundle-branch block; (*d*) arborization (intraventricular) block.

Sinoauricular block. The impulse is originated in the S-A node but does not spread outside the node; therefore the auricular muscle is not stimulated. In the simplest form one or more auricular or ventricular contractions occasionally do not occur; there are "dropped beats." If the block is permanent, an A-V nodal rhythm is established. In the first type the diagnosis can be made by examining the electrocardiogram. The record shows a normal S-A rhythm with an occasional abnormally long diastole, which added to the preceding systole lasts as long as two complete cycles.

If there is complete S-A block, the ECG shows an A-V nodal rhythm, without any signs of S-A activity. The diagnosis of sinoauricular block is difficult to establish, because the A-V nodal rhythm might have been originated in some other way.

FIG. 68. Experimental middle nodal rhythm. ECG (lead II) of a dog in which the S-A node has been destroyed. Direct inspection of the heart showed the auricles and ventricles beating at the same time. There is no P wave; it has been "covered" by the ventricular complex. The arrow points to "splintering" of R. Time in 0.04 sec.

FIG. 69. Record of a human case of middle nodal rhythm. The heart contracts regularly at a rate of 50 beats per minute. There are no P waves; the ventricular complexes are normal.

Auriculoventricular block. This type of block is due to functional depression or anatomical destruction of the bundle of His. Three degrees of block can be produced, experimentally[1] or by disease.

1. *Incomplete A-V block.* The impulses are transmitted with difficulty from the auricle down the bundle of His. There is a prolonged interval between the auricular and ventricular contractions. In the ECG the PR intervals last longer than 0.18 sec. (Fig. 70). In the phlebogram there is an interval of more than 0.2 sec. between waves *a* and *c*.
2. *Partial A-V block.* Some but not all the impulses pass down the bundle of His and stimulate the ventricles. In the ECG some of the P waves are not followed by a ventricular complex (Figs. 42 and 71). In the phlebogram isolated *a* waves can occasionally be recognized (Fig. 42). The cases of partial A-V block are classified according to the ratio between auricular and ventricular beats: 2 : 1, 3 : 1, etc. In some cases there are groups of beats in which the obstacle to the passage of the impulse increases at each successive beat. Thus the PR interval of the first beat is normal (0.18 sec.), that of the second beat is more prolonged, and still more that of the following; finally the fourth P

[1] ERLANGER, J., and A. D. HIRSCHFELDER, *Am. J. Physiol.*, **15**, 153, 1906; **16**, 160, 1906.

wave is not followed by a ventricular complex. The series commences again at the next beat, and so on. These groups (described by the Italian physiologist Luciani) can be made up of a greater or lesser number of beats. The long pause between two series of beats gives time for the complete recovery of excitability, and the following impulse is transmitted at normal speed. This long pause is sometimes called a Wenkebach period.

3. *Complete A-V block.* In this type no impulses are transmitted from the auricles to the ventricles. The auricles beat with one rhythm and the ventricles with another. If the auricles are stimulated by normal S-A impulses, the basal frequency will be 75 to 80 beats per minute, while the ventricles beat with a slower rhythm (idioventricular rhythm) of about 30 beats per minute. In the ECG the P waves are more numerous than the QRST complexes, and there is no definite relation of one to the other, as each has a completely independent origin (Fig. 72). The interpretation of the phlebogram (Fig. 72) presents considerable difficulty in some of these cases.

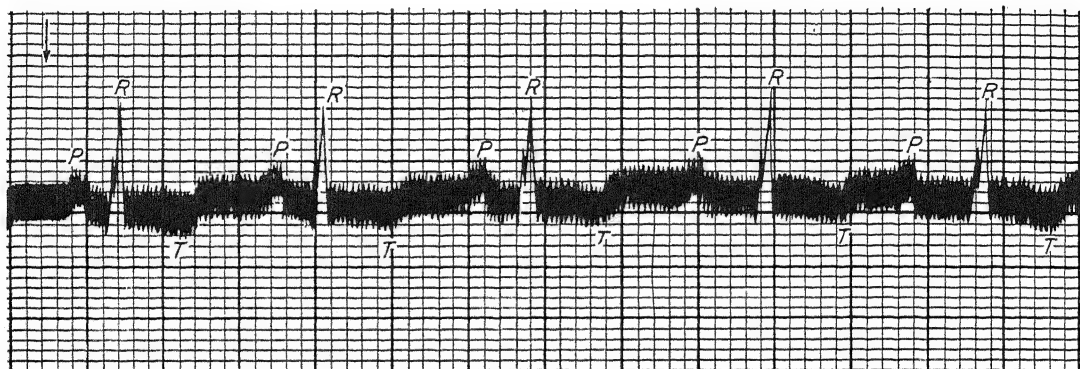



FIG. 70. Incomplete experimental A-V block. Stimulation of the left vagus nerve (beginning at the arrow) prolongs the P-R interval.

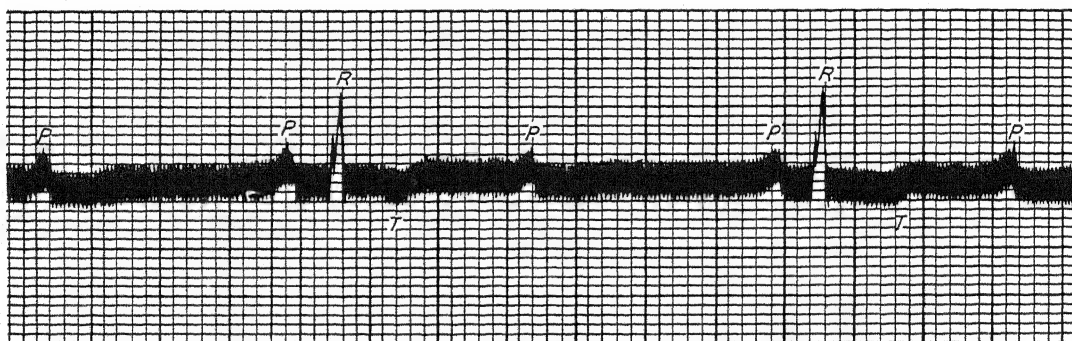



FIG. 71. Partial experimental A-V block. Stimulation of the left vagus nerve with a stronger current than the one used in the experiment recorded in Fig. 70 provokes partial A-V block. Some of the P waves are not followed by ventricular complexes. Every second auricular beat is followed by a ventricular complex. With even stronger stimulation of the vagus it is possible to provoke a complete A-V block.

Bundle-branch block. In this case the obstacle to the spread of the impulse is located in one of the branches of the bundle of His. Excitation of the corresponding ventricle is therefore delayed with respect to that of the ventricle with a normal conduction system. This condition has been produced experimentally[1] and asynchronic contraction of the ventricles has been demonstrated in the intraventricular pressure curves recorded with optical manometers. A

[1] BRAUN MENÉNDEZ, E., and L. A. SOLARI, *Rev. Soc. Argent. de biol.*, **12**, 331, 1936; **13**, 33, 1937.

delay of 0.04 to 0.06 sec. has been observed in the contraction of the ventricle with a severed branch with respect to the contraction of the other ventricle.

The ECG allows the recognition of a bundle-branch block. The rapid part of the ventricular complex is widened and notched. It lasts more than 0.10 sec. and is followed by a large and wide T, usually in the opposite direction to the initial part of the complex. The precordial ECG shows typical changes which permit the diagnosis of the bundle branch which is blocked (see Chap. 15).

Arborization block. There are cases in which the electrocardiogram shows ventricular complexes of little amplitude in all the leads. The initial and final deflections are poorly marked; they are of a vibratory nature, and the initial one lasts more than 0.10 sec. The T wave is usually in the same direction as the initial complex. These cases are interpreted as due to diffuse and extended lesions affecting the Purkinje network.

FIG. 72. Complete A-V block. Phlebogram, phonocardiogram, and ECG in a patient with complete heart block. Time in 0.2 sec. The phonocardiogram shows auricular sounds (*A*) coinciding with *a* waves in the venous pulse and P waves in the ECG. The auricular sound is produced even when the ventricle is in systole and the A-V valves are closed.

PREMATURE CONTRACTIONS AND PAROXYSMAL TACHYCARDIAS

An impulse can be originated in any part of the myocardium where a foreign stimulus acts or where the rhythmicity is enhanced, and thence it can spread over the whole heart. If these abnormal stimuli act occasionally and provoke isolated contractions, the resultant beats are called extrasystoles, or premature contractions, because they anticipate the contraction to be produced by the impulse from the pacemaker that at the time sets the fundamental rhythm of the heart. An extrasystole (or premature contraction) is therefore an anticipated systole provoked by an abnormal impulse. According to the location of their origin they are called "auricular," "A-V nodal" (auriculoventricular), or "ventricular" extrasystoles.

In some cases abnormal stimuli act repeatedly, causing a series of extrasystoles, which may be quite short or may last for some time, producing what is known as paroxysmal tachycardia. The frequency of the heartbeat is considerably above the normal (200, 300, or even more beats per minute). Usually the tachycardia begins and ends suddenly. It can be originated in the auricles, the A-V node, or the ventricles.

The dynamic effects of a premature beat depend on the site of origin of the impulse that originates it and on the path along which it spreads. The later in diastole its occurrence, and the nearer to the normal the path along which

it spreads, the greater will be the dynamic efficiency of a premature beat. An auricular extrasystole which takes place just before the normal systole would have occurred has a dynamic effect not greatly different from a normal systole. On the contrary, an early ventricular extrasystole will have little efficiency; the ventricles will be almost empty and the impulse will be transmitted slowly along abnormal paths.

Auricular extrasystoles. The abnormal impulse may act or can be originated in any part of the auricles. During systole there is complete inexcitability. The nearer to the S-A node and the later in diastole the impulse arises, the more similar will the premature contraction and its consequences be to a normal systole.

As even a normal S-A rhythm may show certain irregularities sometimes, only the ECG will reveal the true nature of the auricular extrasystole. Not only is the beat anticipated; there are also changes in shape, amplitude, duration, and direction of the corresponding P wave. The following ventricular complex usually shows no marked anomaly.

The ventricular diastole following an auricular extrasystole is usually longer than the normal diastole, and it lengthens as the premature contraction occurs later in diastole. Nevertheless the sum of the extrasystolic cycle and the preceding normal cycle is seldom equal to two normal cycles, as is usually the rule in ventricular extrasystoles.

A-V nodal extrasystoles. The abnormal impulses arise in this case in the A-V node. They can be originated in the upper, middle, or lower part of the node. The ECG is that of the corresponding nodal rhythm, previously described.

Ventricular extrasystoles. Ventricular extrasystoles can be produced by external or intrinsic stimuli, acting on any part of the ventricles at any moment of diastole. Except for the rare cases in which the origin is in some part of the bundle of His, the impulse spreads along abnormal pathways. Usually the impulse starts in one of the ventricles; it spreads at low speed, and therefore the ventricles contract asynchronously.

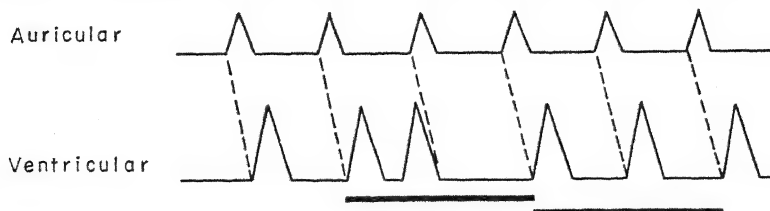
Auricular

Ventricular

FIG. 73. Compensatory pause. Auricular contractions are diagrammatically represented in the upper tracing and ventricular contractions in the lower. The third ventricular beat is premature. The normal impulse from the auricle arrives at the ventricle when it is in systole, and therefore during the refractory period. The ventricle does not respond to this impulse, and contracts only when the following normal impulse arrives. The two consecutive cardiac cycles, which include the premature beat, last as long as two consecutive normal cardiac cycles.

The ventricle where the impulse started contracts first. In some cases the impulse may spread antidromically along the His bundle to the auricles and provoke also an auricular systole.

The dynamics of the heartbeat are disturbed in ventricular extrasystoles much more than in any other type of premature contraction. If the contraction takes place very early in diastole, its efficiency is practically nil. If it occurs somewhat later, blood may be ejected into the arteries, but the contraction is always hypodynamic.

The ECG shows typical changes when ventricular extrasystoles occur. An atypical ventricular complex appears after an interval shorter than normal, with no preceding P wave. The initial part of the complex is widened and notched, lasts more than 0.10 sec., and is followed by a prolonged T wave in the opposite direction to the initial deflection.

The diastole following a ventricular extrasystole is visibly prolonged. It is usually called a compensatory pause. The duration of a normal cycle plus that of the extrasystolic cycle usually lasts as long as two normal cycles. This is due to the fact that ventricular extrasystoles in most cases do not disturb the fundamental S-A rhythm (Fig. 73). When the normal impulse from the pacemaker spreads to the ventricles, it finds them in the systolic refractory state caused by the extrasystole, or having not yet completely

recovered their excitability, and therefore unable to respond. Only the following impulse from the S-A pacemaker will provoke a contraction.

Paroxysmal tachycardias. These are due to a succession of extrasystoles; an ectopic focus commences to discharge impulses at a considerably higher frequency than that of the normal pacemaker. The ectopic rhythm can have its origin in the auricles, the A-V node, or the ventricles. Usually the attack begins and ends suddenly, hence the name of paroxysmal tachycardia (Fig. 74).

The ECG allows the recognition of the origin of the extrasystoles causing the tachycardia, but in some cases even with the help of the ECG it is difficult to make an accurate diagnosis.

Auricular flutter. If the frequency of an auricular tachycardia is 300 or more per minute, the auricles beat regularly but the bundle of His cannot conduct at such a high rate; therefore not all the impulses are transmitted to the ventricles, and a 3:1 to 4:1 A-V block is established. The A-V block may be a total one (Fig. 75). The name of flutter is given to this condition because the flapping of the auricular appendages suggests the fluttering of a bird's wings.

Probably in paroxysmal tachycardia— whatever its origin—and in flutter, processes of circular excitation, such as those discussed when the physiologic properties of the myocardium were considered, come into play. Experimental

FIG. 74. Short paroxysmal tachycardia of ventricular origin, occurring spontaneously in a dog. A series of ventricular extrasystoles, in which there is marked lengthening of the QRS complexes, and deep inverted T waves.

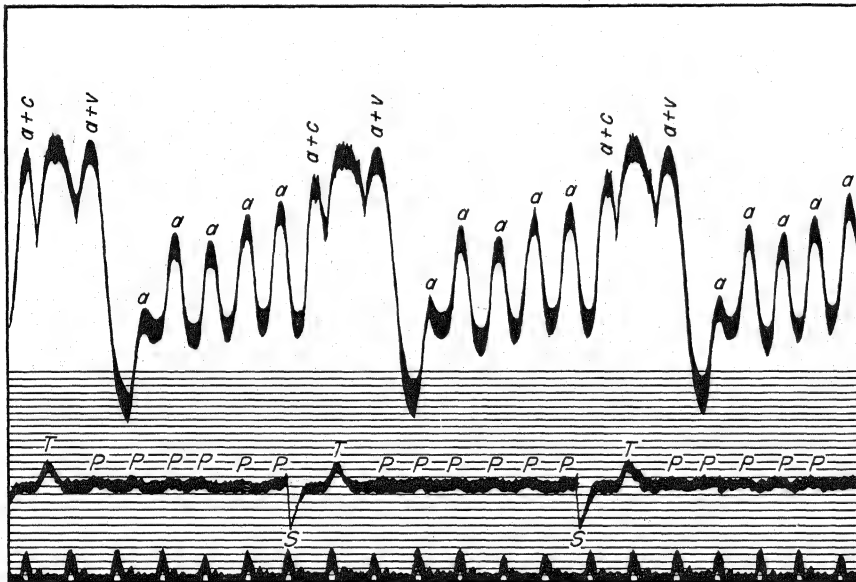



FIG. 75. Experimental auricular flutter, complete A-V block, and bundle-branch block. Phlebogram, ECG, and time in 0.2 sec. Rough handling of the auricles caused auricular flutter in a dog in which the bundle of His had been cut. The ECG also shows there is a bundle-branch block. (*Courtesy of Dr. A. Caeiro.*)

damage of the auricular wall in the intervenous area facilitates the development of auricular flutter in the dog; the frequency of auricular beats is directly proportional to the size of the area destroyed.[1]

Incoordinated Excitation. Auricular and Ventricular Fibrillation

In all the cases considered so far, the process of excitation is produced in an orderly way, and a coordinated contraction of the heart muscle results. The excitability of the heart is such that in certain conditions a disorderly process of excitation can be established, characterized by the simultaneous presence of fibers in contraction and in relaxation. There is a complicated and constantly changing network of active fibers interwoven into a no less complicated and changeable network of quiescent fibers. The condition is called fibrillation. When the muscular mass is sufficiently great, fibrillation tends to become permanent. Fibrillation can occur in the auricles, in the ventricles, or in both at the same time.

Auricular fibrillation is compatible with life and even with excellent circulatory dynamics. The auricular muscle does not contract coordinately; a fine, irregular tremor moves all its fibers. The efficiency of this type of activity is nil, and the auricles do not propel the blood forward. The ventricular systole is coordinated, but ventricular beats are separated by extremely irregular intervals. Their vigor is also irregular. No two beats are equally vigorous or equally spaced. There is a complete or permanent arrhythmia (delirium cordis). Nevertheless circulatory conditions may be quite satisfactory.

Auricular fibrillation can be provoked in dogs by stimulating the auricles with a tetanizing current. It is usually of short duration; after a few minutes it is gradually transformed into flutter, and then suddenly the normal S-A rhythm is reestablished. There are no P waves in the ECG when there is auricular fibrillation. The ventricular complexes are normal but they are spaced irregularly (Fig. 87). During diastole an irregular vibration of small amplitude is seen in the record, and occasionally a small wave appears which is not sufficiently large to be identified with a P wave. This persisting electrical activity is due to the activity of the auricular fibers.

[1] Rosenblueth, A., and J. García-Ramos, *Acta cardiol.*, Suppl. 2, 87, 1947.

Ventricular fibrillation is nearly always permanent and a relatively common final event, as far as life is concerned, in large animals and in man. Blood is not ejected from the heart and the arterial blood pressure falls to zero. The ventricles become considerably distended by the blood which continues to return to the heart. The large veins are filled with blood.

Several causes may lead to ventricular fibrillation: electric currents of all types, mechanical stimulation, drugs, etc. A frequent cause of ventricular fibrillation is the occlusion of an important branch of a coronary artery.

Ventricular fibrillation has been the most serious immediate complication in a few cases of intracardiac catheterization and surgical operations on the heart. Most of these patients died. When these procedures are carried out, care should be taken to prevent ventricular fibrillation and to have at hand means of controlling it should it occur. In order to prevent fibrillation the heart should not be handled roughly, scratched, or pinched; the endocardial surface is especially sensitive. Certain drugs have an antifibrillating action, *e.g.*, novocainamide.[1] If fibrillation does occur, it is still possible to reestablish coordinated beats, but unfortunately the procedure to do this is not easy to apply in the conditions in which fibrillation usually occurs. The most efficacious treatment consists in the application of an electric countershock: two relatively large electrodes are placed directly on the surface of the exposed heart and a 60-cycle, 5-amp. current of very short duration is passed.[2] While the apparatus is being prepared a certain degree of blood pressure must be maintained by rhythmic compression of the ventricles (heart massage). Electric countershock has been applied successfully in human cases of ventricular fibrillation.[3]

The ECG in cases of ventricular fibrillation shows only a completely irregular and dis-

[1] Wedd, A. M., H. A. Blair, and R. S. Warner, *Am. Heart J.*, **42**, 399, 1951; Kayden, H. J., J. M. Steele, L. C. Mark, and B. B. Brodie, *Circulation*, **4**, 13, 1951; Harris, S., A. Estandia, T. J. Ford, H. T. Smith, R. W. Olsen, and R. F. Tillotson, *Circulation*, **5**, 551, 1952.

[2] Stearns, N. S., G. L. Maison, and J. W. Stutzman, *Am. J. Physiol.*, **164**, 601, 1951.

[3] Southworth, J. L., V. A. McKusick, E. C. Peirse. and F. L. Rawson, *J. A. M. A.*, **143**, 717, 1950.

orderly series of deflections of unequal size and low voltage.

Ventricular fibrillation consists at first of a fairly vigorous twitching of the myocardiac fibers. Gradually these become weaker, and after 10 to 20 min. their activity is reduced to a few weak waves which slowly travel along the heart muscle. When they finally cease the heart muscle is dead.

BIBLIOGRAPHY

BATTRO, A., "Arritmias," 2d ed., El Ateneo, Buenos Aires, 1935.

COSSIO, P., "Aparato Circulatorio," El Ateneo, Buenos Aires, 1944.

GLOMSET, D. J., and A. T. A. GLOMSET, A Morphologic Study of the Cardiac Conduction System in Ungulates, Dog, and Man, *Am. Heart J.*, **20**, 389 and 677, 1940.

LEWIS, T., "Mechanism and Graphic Registration of the Heart Beat," 3d ed., Shaw and Sons, London, 1925.

LUISADA, A. A., "Heart, A Physiological and Clinical Study of Cardiovascular Disease," Williams & Wilkins, Baltimore, 1948.

MANHAIM, I., "Les Maladies organiques du faisceau de His-Tawara," Masson et Cie, Paris, 1931.

ORÍAS, O., "Registro e Interpretación de la Actividad Cardíaca," 3d ed., El Ateneo, Buenos Aires, 1939.

PRINZMETAL, M., E. CORDAY, I. C. BRILL, R. W. OBLATH, and H. E. KRUGER, "The Auricular Arrhythmias," Charles C Thomas, Springfield, Ill., 1952.

RIJLANT, P., "La Conduction dans le coeur du mammifère," Vaillant-Carmanne, Liège, 1931.

ROTHBERGER, C. J., in Bethe's "Handbuch der normalen und pathologischen Physiologie," **7**, 523, Springer, Berlin, 1926.

ROTHBERGER, C. J., *Ergebn. d. Physiol.*, **32**, 472, 1931.

TIGERSTEDT, R., "Physiologie des Kreislaufes," Walter de Gruyter & Company, Berlin and Leipzig, 1921.

WENKEBACH, K. F., and H. WINTERBERG, "Die unregelmässige Herztätigkeit," 2d ed., W. Engelmann, Leipzig, 1927.

WHITE, P. D., "Heart Disease," 3d ed., Macmillan, New York, 1944.

CHAPTER 17

The Nervous Control of Cardiac Activity

THE HEART HAS within itself all the factors necessary to originate and maintain its activity provided it is in an adequate medium. Rhythmicity and the other physiologic properties are inherent in the myocardial fiber, independently of the nervous tissue included in the heart. Nevertheless the nervous system plays a part in coordinating and integrating the activity of the heart with the needs of the organism in different physiologic circumstances.

The nervous regulation of the heart is accomplished through the autonomic nervous system (Chap. 84). From a functional point of view three systems may be considered in so far as the nervous regulation of the heart's activity is concerned: (*a*) inhibitory; (*b*) accelerator; (*c*) sensory. The conception of an intracardiac, locally integrated nervous system is no longer tenable. Intracardiac ganglia, at any rate in vertebrates, are only relay stations along the efferent paths.

THE CARDIOINHIBITORY SYSTEM

In 1845 E. H. and E. F. Weber[1] demonstrated that stimulation of the vagus nerve in the frog stops the heart in diastole. Up to that time only motor effects had been observed on peripheral nerve stimulation; for the first time an inhibitory effect was obtained. It is now common knowledge that inhibition, although less conspicuous than excitation, is not less important nor less frequent in the physiology of the nervous system. Soon after this pioneer demonstration, the inhibitory effect of the vagus on the heart was observed in many other animal species and the principle was extended to all the vertebrates.

The vagus nerve is part of the cranial parasympathetic division of the autonomic nervous system. The cellular bodies of its efferent fibers are situated in the medulla, in what is known in mammals as the dorsal nucleus. The greater part of the fibers of the right vagus end in the S-A node, the remainder in the A-V node. The greater part of the fibers of the left vagus end in the A-V node, and only a few in the S-A node. For this reason stimulation of the right vagus has a greater effect on the rhythm of the heart as a whole, whereas stimulation of the left vagus usually produces heart block (Figs. 70 and 71). The fibers of the vagus end on ganglion cells, which send out postganglionic fibers to the auricles and the basal portion of the ventricles.

Analysis of the vagal action on the heart. The effects of stimulation of the vagus in the frog, toad, tortoise, dog, cat, and rabbit (animals most frequently used in laboratory experiments) are essentially the same in all species. The animal is anesthetized, and the cervical part of the vagus is dissected; it is cut so that stimulation will not spread to the nerve centers along the afferent fibers and thus provoke unwanted reflex effects. The peripheral end is stimulated, using bipolar electrodes and a tetanizing current from an induction coil. The effects produced can be observed directly on the exposed heart, or indirectly by recording the variations in blood pressure, arterial pulse, the electrocardiogram, or any other aspect of cardiac activity. There is always a short latent period between the application of the stimulus and its effect.

A stimulus of moderate intensity slows the

[1] WEBER, E. H., *Ann. universali di medicina*, **116**, 225, 1845.

heart rate (bradycardia). A strong stimulus stops the heart completely. If the left vagus is stimulated with moderate intensity, complete or partial A-V block is almost always provoked (Figs. 70 and 71). Strong stimulation of the left vagus also brings the heart to a standstill. The arterial blood pressure falls rapidly and reaches zero if the heart is stopped for a sufficient time (Fig. 76, 1). The arterial pulse vanishes and electrical variations connected with the heart's activity disappear completely. (Usually the stimulating current spreads through the body to the electrodes, and may be recorded by the electrocardiograph.)

As soon as vagal stimulation ceases, the heart renews its activity and usually a period of tachycardia follows. The blood pressure rises rapidly to the initial level and for a short time may surpass it. The arterial pulse and the electrical signs of cardiac activity reappear.

Stimulation of the vagus nerve depresses not only the rhythmicity of the heart but also its conductivity and contractility. On the other hand, excitability is not modified; the threshold remains unaltered during stimulation of the vagus. The refractory period, however, is shortened and the threshold of fibrillation is lowered during the vulnerable period.[1] Stimulation thresholds also remain unaltered while the heart is stopped by the effect of acetylcholine.[1]

Vagal action is exerted on the nodal tissues and on the auricular myocardium. The ventricles cease beating because no impulses reach them. When there is a complete A-V block, vagal stimulation does not have any effect on the idioventricular rhythm.

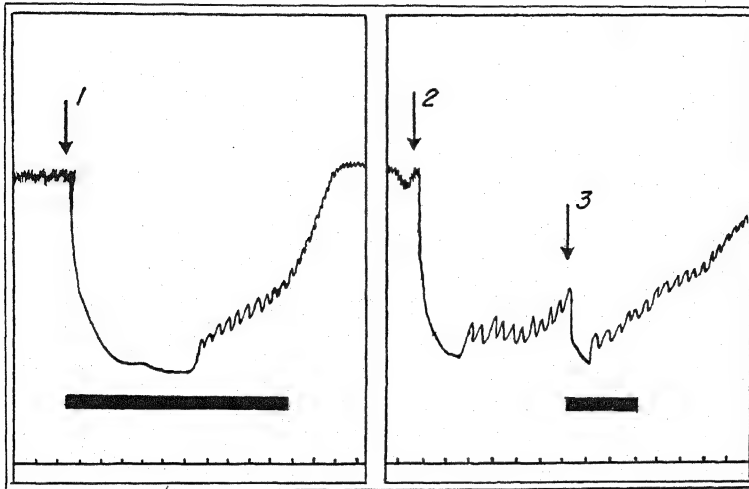



FIG. 76. Vagal action and vagus "escape." Arterial blood-pressure curves registered by a mercury manometer. 1, stimulation of the peripheral end of the right vagus nerve; the heartbeat stops for a time, but begins again in spite of the continuous stimulation (black line) of the vagus. 2, injection of acetylcholine. When the heart is again beating, at 3, the right vagus is stimulated as in 1; "escape" occurs much sooner. Time in 5 sec. (*Diagram of an experiment by A. S. Segura, Rev. Soc. argent. de biol., vol. 14, p. 497, 1938.*)

Mechanism of vagal action. When a heart is perfused and the vagus stimulated, if the perfusion fluid is used to perfuse another heart, it will produce bradycardia in this second heart or stop it, as if the vagus had been stimulated. This fact was interpreted as proof that stimulation of the vagus sets free a chemical substance which diffuses into the perfusion fluid, where it can be demonstrated by its pharmacological effects.[2] This substance, first called the "vagal substance," is acetylcholine. This is a particular instance of the release by nerve stimulation of a chemical mediator for the transmission of the impulses at the synapse and from nerve fibers to effectors (see Chap. 84). Acetylcholine is the chemical mediator in the parasympathetic division of the autonomic nervous system; both the preganglionic nerve endings in the ganglia

[1] HOFFMAN, B. F., A. A. SIEBENS, and C. McC. BROOKS, *Am. J. Physiol.*, **167**, 796, 1951.

[1] MOISSET DE ESPANÉS, E., *Rev. Soc. argent. de biol.*, **26**, 257, 1951; *Compt. rend. soc. de biol.*, **145**, 452, 1951.

[2] LOEWI, O., *Pflüger's Arch. f. d. ges. Physiol.*, **189**, 239, 1921.

and the postganglionic nerve endings on the effectors liberate acetylcholine. The cardioinhibitory system belongs to the cranial division of the parasympathetic.

Intravenous injection of acetylcholine produces bradycardia or stops the heart, according to the dose given. It has the same effect as stimulation of the vagus (Fig. 76, 2). The auricles of a rabbit which have ceased to beat after 24 to 36 hr. of perfusion with Tyrode's solution will begin to beat again if a small quantity of acetylcholine (1:100,000,000 to 1:400,000,000) is added to the perfusion fluid. When the heart has thus recovered its activity, a larger dose of acetylcholine will again stop it.[1] Atropine, an alkaloid extracted from the plant *Atropa belladonna* (deadly nightshade), suppresses vagal inhibition of the heart and the effects of acetylcholine.

During stimulation of the vagus, potassium is also liberated, and as potassium has a depressor activity on the heart, at one time it was thought to be the substance responsible for the action of the vagus.[2] The effect of potassium is not identical to that of stimulation of the vagus, since it is not suppressed by atropine. Probably there is some connection between the liberation of acetylcholine and that of potassium, but its nature has not yet been established.

Muscarine, pilocarpine, and eserine have an effect similar to that produced by stimulation of the vagus, but none of these substances is as perfectly parasympathicomimetic as acetylcholine.

"Vagus escape." If stimulation of the vagus is continued for some time after the heart has stopped, the heart again begins to beat in spite of the persistent stimulation. The heart "escapes" from the influence of the vagus, hence the name of "vagus escape." There are two possible explanations of this phenomenon: either acetylcholine is exhausted by prolonged stimulation, or the heart becomes refractory to the effect of acetylcholine. It is not yet possible to deny the former explanation, but there is proof that acetylcholine soon loses its inhibitory effect on the heart. A sudden increase in the concentration of acetylcholine will again produce cardiac inhibition, but this will also be of a transitory nature even though the concentration of acetylcholine is maintained constantly at the new high level.[1] Acetylcholine behaves, with respect to its cardioinhibitory activity, as a "potential drug."[2] Moreover, stimulation of the vagus after injection of acetylcholine produces a much less marked effect than vagal stimulation without the previous injection of this drug (Fig. 76). These experiments are arguments in support of the hypothesis that "vagus escape" is due to a kind of tolerance of the myocardiac fiber to acetylcholine.

[1] Bülbring, E., and J. H. Burn, *J. Physiol.*, **108**, 508, 1949.

[2] Howell, W. M., *Am. J. Physiol.*, **15**, 280, 1906; **21**, 51, 1908.

Reflex Cardiac Inhibition

Direct stimulation of the vagus, such as has been considered so far, seldom occurs outside experimental conditions. In physiologic conditions the cardioinhibitory system is stimulated by reflex mechanisms. Some of these mechanisms are activated accidentally and may produce harmful effects. Others are permanent physiologic processes which take part in the regulation of fundamental equilibriums, such as the arterial blood pressure.

Accidental cardioinhibitory reflexes can be provoked by stimulation of any sensory nerve, but from some particular areas marked vagal effects on the heart can be produced. Thus sudden or gradual abdominal compression provokes bradycardia, or even stops the heart (Goltz's reflex) for a sufficient time to cause fainting and unconsciousness (*e.g.*, a knockout blow on the epigastrium). The nasal mucosa when stimulated by chemical irritants, such as formaldehyde, ammonia, chloroform, etc., initiates cardioinhibitory reflexes that are sometimes the cause of serious accidents in the course of anesthesia. Compression of the eyeballs produces bradycardia in subjects with a well-balanced autonomic nervous system (Aschner-Dagnini phenomenon). In Ortner's phenomenon (forced extension of the neck) and Czermak's phenomenon (compression of the carotid at the level of the thyroid cartilage), bradycardia is produced reflexly by changes in pressure in the carotid sinus and not by direct stimulation of the vagus nerve as formerly believed. A sufficient

[1] Moisset de Espanés, E., and C. Martínez, *Rev. Soc. argent. de biol.*, **14**, 528, 1938.

[2] Drugs with "potential action" are those that produce a visible effect only when there is a sudden increase in their concentration. If the concentration remains constant or increases very gradually, they have no effect.

dose of atropine suppresses all these vagal reflexes.

Pressoreceptor areas in the aorta and carotid sinus. Nerves of de Cyon and Hering. The inner surface of the first part of the aorta and the carotid sinus (a dilatation at the beginning of the internal carotid) are the sites of origin of reflex impulses that exert a permanent influence on the cardioinhibitory system (Figs. 77 and 78). Nerve impulses originated in these areas, in special receptors sensitive to pressure changes, are conveyed to the medulla by the cardioaortic and sinus nerves. The cardioaortic nerves or the nerves of de Cyon are formed by fibers arising mainly in the internal layers of the aorta and also to a lesser extent in the endocardial layers of the left ventricle. The fibers arising in the left ventricle and the root of the aorta form the right cardioaortic nerve; the fibers arising in the aortic arch form the left cardioaortic nerve. In man and in the dog the cardioaortic nerves are incorporated in the vagi; in the rabbit they constitute separate nerves. Stimulation of the peripheral end of the cardioaortic nerves, after having cut them in the neck, produces no effect; stimulation of the proximal end, on the contrary, is followed by bradycardia and a fall in blood pressure. For this reason the aortic nerves were called "inhibitory nerves."[1]

FIG. 77. Innervation of the vasosensory areas. 1, heart; 2, aorta; 3, common carotid artery; 4, carotid sinus; 5, external carotid; 6, internal carotid; 7, carotid body; 8, de Cyon's cardioaortic nerve; 9, vagal ganglion; 10, nerve fibers from the carotid sinus to the glossopharyngeal nerve; 11, nerve fibers from the carotid sinus to the vagus nerve; 12, glossopharyngeal nerve; 13, vagus nerve. (*C. Heymans.*)

The fibers of the sinus nerves, also called "nerves of Hering," one on each side of the body, arise in the internal layers of the walls of the corresponding sinus caroticus and immediately join the glossopharyngeal nerve on the respective side. Anatomically the sinus nerves are inconspicuous, and they were not noticed until Hering (1924) drew attention to their functional significance. Afferent stimulation of these nerves is followed by bradycardia and a fall in arterial blood pressure.

FIG. 78. Position of the carotid sinus region in man

The cardioaortic and sinus nerves are sometimes called "pressoreceptor" or "vasosensory" nerves. By recording the action currents of these nerves it has been possible to demonstrate that

[1] DE CYON, E. and K. LUDWIG, *Ber. Sächs. Ges.*. 18. 307, 1866.

when the blood pressure is normal, impulses of low frequency are discharged along their fibers, especially at each systolic rise of pressure. When the blood pressure rises, the frequency of the impulses increases and even during the diastolic fall in pressure the discharge persists (Fig. 79).

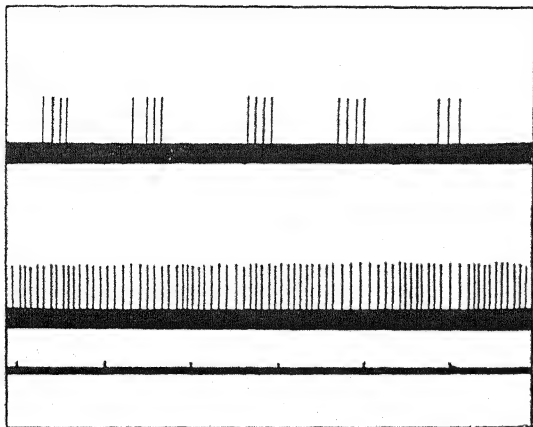



FIG. 79. Action currents of a pressoreceptor. Above: nerve impulses from an isolated receptor in the carotid sinus when the arterial pressure was 55 mm. Hg; each series of impulses was provoked by the rise in pressure at each systole. Below: the arterial pressure was raised to 135 mm. Hg; the frequency of the impulses increased considerably, and the discharges continued during diastole. (*Diagram from an experiment by Bronk and Stella, J. Cell. & Comp. Physiol., vol. 1, p. 113, 1932.*)

On the contrary, when the blood pressure falls, the frequency of the impulses diminishes, and when the pressure is very low the impulses cease completely.

The impulses travel along the depressor nerves to the cardioinhibitory center and, through the vagus, cause a decrease in the frequency of the heartbeat. This reflex controls the blood pressure; an increase in blood pressure diminishes the heart rate, and thus the blood pressure is lowered. The effect of these nerves is not limited to the heart rate; the impulses spread to the vasomotor center, and arterial vasodilatation is provoked, which also lowers the blood pressure.

Marey[1] pointed out many years ago the inverse relation existing between the blood pressure and the heart rate (Marey's law).

Changes in the internal pressure of the arteries are not the only efficient stimulus of the pressoreceptors. Compression with a forceps or the fingers, or elongation of the artery by stretching it, also produces a reflex response.

Section of all the cardioaortic and sinus nerves is followed by tachycardia and hypertension, which is permanent in some animals; in the dog, pulse rate and blood pressure gradually return to normal.

Carotid and aortic bodies.[1] At the bifurcation of the common carotid artery, between the internal and external carotid arteries, there is a small structure known as the carotid body. It is made up mainly of vascular sinusoids, the walls of which are of endothelial nature; special cells and a rich nerve plexus complete the structure of this peculiar little organ. The nerves arising in the carotid bodies (one on each side) are not sensitive to distention, but they are stimulated by changes in the chemical composition of the blood, eliciting cardiovascular and respiratory responses. The carotid bodies can be considered as a complement to the carotid sinus. The aortic body situated near the arch of the aorta is similar in structure and function to the carotid bodies. Stimulation of the carotid and aortic bodies by chemical changes in the blood is not of much importance in the regulation of cardiac activity and blood pressure, but it is of great significance in the regulation of respiration (Chap. 32).

Vagal tone. Normal arterial blood pressure stimulates the vascular pressoreceptors, and the impulses from these receptors are constantly stimulating the cardioinhibitory center, which discharges along the vagi nerves. The heart is thus submitted to a continuous inhibitory influence. If both vagi are cut, or atropine is injected into an animal with normal blood pressure, the heart rate increases immediately. The permanent influence of the vagi on the heart is called vagal tone. A decrease of vagal tone is the first step to cardiac acceleration. Respiratory fluctuations in vagal tone determine periodic changes in the heart rate, which increases during inspiration and decreases during expiration (respiratory sinus arrythmia). Distention of the right auricle by an increased venous return also provokes tachycardia (Bainbridge's reflex).[2]

Cardioinhibitory center. The center for the cardioinhibitory reflexes is situated in the medulla. It is composed of cellular bodies from

[1] MAREY, E. J., "Physiologie médicale de la circulation du sang," Paris, 1863, p. 206.

[1] SCHMIDT, C. F., and J. H. COMROE, *Physiol. Rev.*, 20, 115, 1940.

[2] BAINBRIDGE, F. A., *J. Physiol.*, 50, 65, 1915.

which the efferent fibers of the vagus arise, (dorsal nucleus of the vagus) and also of the portions of gray substance where the afferent fibers of the vagus end. The vasomotor center is very close to the cardioinhibitory center, and both centers may be influenced by the same factors. The cerebral cortex may also send impulses to the cardioinhibitory center and modify its activity, as in changes in heart rate due to emotional states. The destruction of the cardioinhibitory center, or its suppression by drugs, abolishes vagal tone and all the cardioinhibitory reflexes.

Trophic action of the cardioinhibitory system. The vagus has a definite effect on cardiac metabolism, favoring the preservation, and recovery through anabolic restoration, of substances that liberate energy on disintegration. In this respect as well, the vagus is antagonistic to the sympathetic.

THE CARDIOACCELERATOR SYSTEM

E. and M. de Cyon in 1867 gave a conclusive demonstration of the existence of nerves that, when stimulated, produced acceleration of cardiac activity. They pointed out, at the same time, the course followed by the fibers.[1] These nerves belong to the thoracolumbar division of the autonomic nervous system. The path of cardioaccelerator fibers has been determined in man by studying the results of therapeutic cervicothoracic sympathectomy. The following conclusions are now well established:[2] (*a*) few, if any, fibers emerge from the spinal cord in the first thoracic nerve; (*b*) the greater part of the fibers form part of the second thoracic nerve; (*c*) important numbers of fibers emerge with the third, fourth, and fifth thoracic nerves; (*d*) the fifth thoracic nerve seems to be the lowest level at which cardioaccelerator fibers leave the spinal cord; (*e*) more fibers emerge on the right than on the left side. This distribution does not vary significantly from that found in other mammals (Fig. 80). The most important group of accelerator fibers passes through the annulus of Vieussens. At this point some of the fibers are still preganglionic; they end on neurons in the inferior and middle cervical ganglia, but most of them are postganglionic fibers that have their cellular bodies in the thoracic ganglia. From the annulus of Vieussens and from the cervical ganglia, the accelerator fibers proceed in the cardiac nerves and end in the heart. Smaller groups of postganglionic fibers go directly from the upper thoracic ganglia to the heart.

Direct stimulation of the cardioaccelerator nerves. The most favorable place for stimulation of the cardioaccelerators is the stellate ganglion, which can be reached from the upper dorsal region without opening the pleura, or after having opened the thorax in an anesthetized animal kept alive by artificial respiration.

The effect of direct stimulation of the accelerator nerves is not so dramatic as the effect of stimulating the vagi. The latent period is longer (it varies between 10 and 20 sec.) and the heart rate increases gradually. The increase in heart rate is greater when the initial frequency is low than when there is already some degree of tachycardia, and it persists for some time after stimulation has ceased. The operation needed to reach the stellate ganglion causes tachycardia, and for this reason the result of experimental stimulation of the accelerators is also less marked than that of vagal stimulation.

Cardiac acceleration is due principally to shortening of the ventricular diastole, but the ventricular systole is also shortened, as Wiggers has shown. This shortening of the systole is more marked in cardioacceleration caused by stimulation of the sympathetic nerves than in tachycardia due to other causes. There is also considerable shortening of systole in adrenaline tachycardia.

The cardioaccelerator effect of the sympathetic is known as a positive chronotropic effect. The other fundamental properties of the myocardium (excitability, conductivity, and contractility) are also increased by sympathetic stimulation. With respect to excitability, only a presumptive conclusion can be arrived at, because adequate and sufficiently accurate methods have not yet been applied in the study of this problem. Sympathetic stimulation favors fibrillation of the heart.

The blood pressure rises when the accelerators are stimulated. This effect is due to the increase in heart rate. The frequency never increases sufficiently to impair the filling of the heart.

The sympathetic system exerts, to a certain extent, a direct effect on the ventricles. In cases of complete A-V block, stimulation of the

[1] De Cyon, E., and M. de Cyon, *Arch. f. Anat. u. Physiol.*, p. 369, 1867.

[2] Smithwick, R. H., E. M. Chapman, D. Kinsey, and G. P. Whitelaw, *Surgery*, **26**, 727, 1949.

accelerator nerves increases the frequency of both auricular and ventricular contractions, but the rhythm of each chamber remains independent of the other.

other adrenergic sympathetic fibers are stimulated (see Chap. 84). According to von Euler,[1] sympathin is made up of adrenaline and noradrenaline in variable proportions. Sympathin

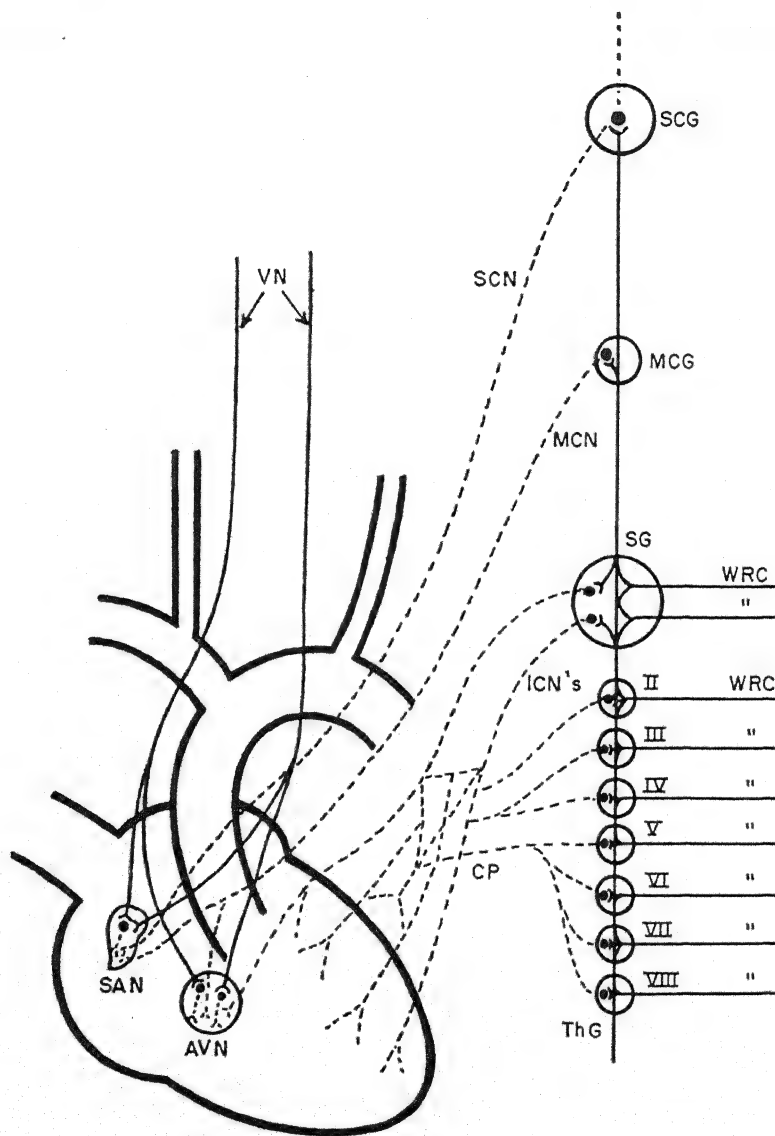



FIG. 80. Efferent innervation of the heart. *VN*, vagus nerve; *SCG*, superior cervical ganglion; *MCG*, middle cervical ganglion; *SG*, stellate ganglion; II to VIII *ThG*, upper thoracic ganglia; *WRC*, white rami communicantes; *SCN*, superior cardiac nerve; *MCN*, middle cardiac nerve; *ICN's*, inferior cardiac nerves; *CP*, cardiac plexus; *SAN*, sino-auricular node; *AVN*, auriculoventricular node.

Mechanism of cardioaccelerator action. A chemical mediator—sympathin[1]—is released at the peripheral endings of the accelerator fibers when they are stimulated, just as when any

[1] CANNON, W. B., *et al., Am. J. Physiol.*, **96**, 392, 1931; **97**, 365, 1931; **99**, 398, 1932.

extracted from human, bovine, and porcine hearts is very similar to noradrenaline.[2] Noradrenaline injected intravenously has similar effects on the

[1] VON EULER, U. S., *Ergebn. d. Physiol.*, **46**, 261, 1950.
[2] RABB, W., and E. LEPESCHKIN, *Circulation*, **1**, 741, 1950.

heart to those of adrenaline, but they are less marked. In man it provokes a sharp rise in systolic and diastolic blood pressure, which causes reflex bradycardia. Atropine abolishes this bradycardia; a cardioaccelerator effect of noradrenaline can be observed only if atropine has been previously injected.

Other drugs of similar chemical structure to that of adrenaline, such as ephedrine, tyramine, veritol, etc., have an effect very similar to that of adrenaline. Nicotine also increases the heart rate. Ergotamine, an alkaloid extracted from the ergot of rye, suppresses the accelerator effect of both sympathetic stimulation and adrenaline injection.

Cardioaccelerator tone. Bilateral removal of the stellate ganglia causes a permanent decrease in the heart rate. This fact has been interpreted as evidence of a permanent "tonic" action of the cardiac accelerator nerves, which is usually masked by the more powerful vagal tone.

Cardioaccelerator center. From a functional standpoint a cardioaccelerator center may be admitted; it is situated near the cardiodepressor center but is not so clearly differentiated. When the cardioaccelerator center enters into activity it provokes a more marked tachycardia than that caused by mere inhibition of the cardiodepressor center. Agents that at first cause inhibition of the cardiodepressor center stimulate the cardioaccelerator center when the intensity of their action increases. This occurs when there is a marked fall in blood pressure, a great distention of the auricles, an intense emotional state, etc.

Trophic action of the cardiac sympathetic. Stimulation of the cardiac sympathetic activates catabolic processes, increasing the consumption and disintegration of substances that thereupon release energy. Oxygen consumption also increases.

THE SENSORY INNERVATION OF THE HEART

The afferent fibers distributed in different parts of the heart provide this organ with a particular form of sensibility, which enters consciousness only in certain conditions. The normal functioning of the heart remains completely outside the sphere of consciousness. Only when the heartbeat increases in strength and frequency, under the influence of exercise, emotions, or stimulating drugs, is it felt as "palpitations." Premature contractions (extrasystoles), especially those of ventricular origin, are also perceived as a sudden and fleeting sensation of constriction, vaguely localized in the cervical part of the respiratory tract, or in the epigastrium.

The heart has no tactile sensibility, as was demonstrated by Harvey in the historic case of the Earl of Montgomery, who had an exposed heart due to thoracic malformation.

From a medical point of view the most important form of sensibility is that of pain, which appears in cases of myocardiac ischemia due to diffuse or localized disturbances in the coronary circulation. The patient experiences a sensation of suffocating contraction localized deeply but diffusely within the chest. There is also a most intense and acute pain referred to the precordial region and the left side of the chest, irradiated along the left arm down to the hand and the fourth and fifth fingers; an agonizing sensation of imminent death completes the picture. This type of pain is known as anginal pain, or angina pectoris.

The course of the nerve fibers along which the impulses responsible for these painful sensations travel has been carefully investigated, mainly with the object of devising surgical procedures for relieving patients in whom the pain has become intolerable because of its intensity and persistence. The nerve fibers leave the cardiac plexus in the middle and inferior cardiac nerves and go to the middle cervical and stellate ganglia on the left side of the body. They pass through these ganglia and down the left sympathetic chain, which they leave in the rami of the first four or five thoracic nerves, penetrating into the spinal cord by the corresponding dorsal roots. Other fibers go directly from the plexus to the first five or seven thoracic nerves. Finally a few fibers join the vagus nerve (Fig. 81). Surgical procedures for the relief of cardiac pain consist in the extirpation of the stellate ganglion on the left side of the body, which interrupts the principal path, or in cutting or destroying with an injection of alcohol, or anesthetizing, the five upper thoracic rami or the dorsal roots of the corresponding thoracic nerves.

Precordial pain and its irradiation along the left arm belong to the type known as referred pain. This type of pain is felt in the somatic structures innervated by the same segment as

the viscera in which the pain has its origin (see Chap. 73).

Inflammation of the pericardium and pericardial effusion also provoke precordial pain irradiated to the neck or the epigastrium.

myocardial fibers. The following are some of the most conclusive demonstrations:

1. When the intracardiac ganglia of the frog's or turtle's heart are extirpated, the organ not only does not stop, but no fundamental change is observed in the sequence of the heartbeats under the new circumstances.
2. Suppression of the activity of the intracardiac ganglia by treating them with nicotine has no disturbing effect on the heartbeat.
3. Small fragments of heart muscle, completely free from nerve cells, contract rhythmically.

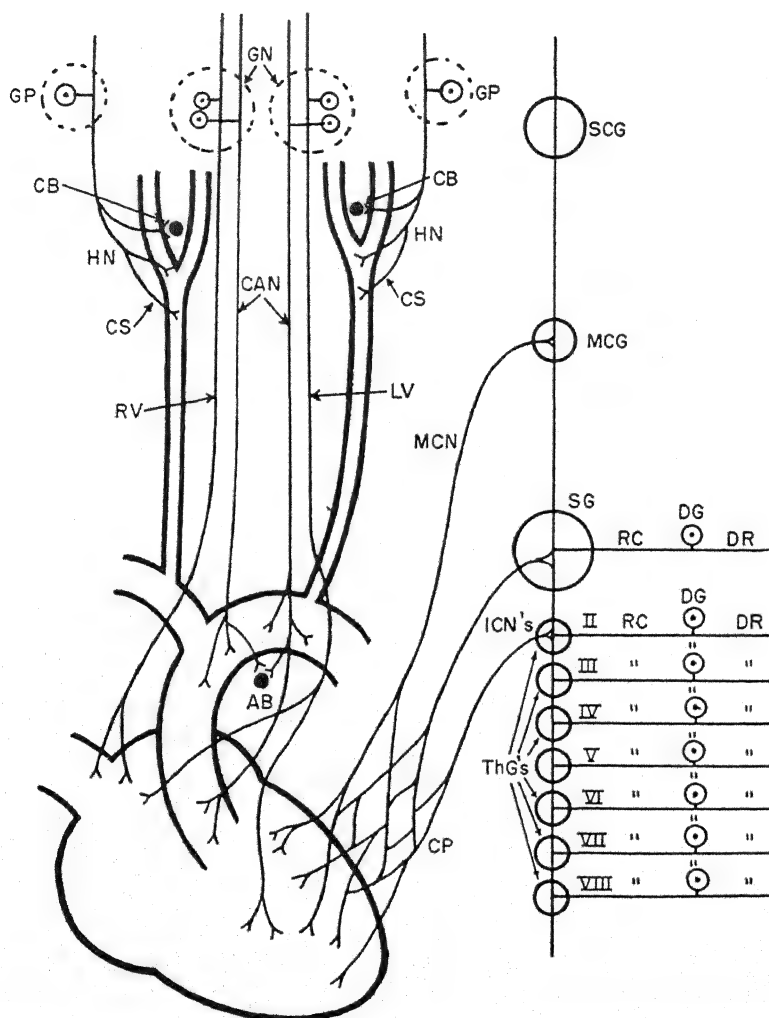



FIG. 81. Afferent innervation of the heart. *GP*, ganglion petrosum; *GN*, ganglion nodosum; *SCG*, superior cervical ganglion; *MCG*, middle cervical ganglion; *SG*, stellate ganglion; II to VIII *ThG's*, upper thoracic ganglia; *RC*, rami communicantes; *DG*, dorsal ganglia; *DR*, dorsal roots; *CB*, carotid body; *HN*, Hering's nerve; *CS*, carotid sinus; *CAN*, cardioaortic nerve; *RV*, right vagus nerve; *LV*, left vagus nerve; *AB*, aortic body; *MCN*, middle cardiac nerve; *ICN's*, inferior cardiac nerves; *CP*, cardiac plexus.

THE MYOGENIC OR NEUROGENIC ORIGIN OF CARDIAC ACTIVITY

At one time the nervous or muscular origin of the impulse that causes the heartbeat was the object of considerable debate. The matter seems now to have been settled. In vertebrates there is definite proof that the impulse starts within the

4. In the chick embryo the heart begins to beat before any nervous structure is present.
5. Tissue cultures of myocardium, in which only cardiac muscle fibers are growing, contract rhythmically.

The capacity to conduct impulses is also a property of the myocardial fibers.

The structure and the properties of the heart in the invertebrates vary considerably from one species to another, so it is difficult to arrive at general conclusions valid for all. Thus the heart of snails of the genus *Helix* is made up of plain muscle, while that of the crustacean *Limulus polyphemus* (the horseshoe crab) is constituted of ordinary striated-muscle fibers. Carlson's experiments[1] have shown that in *Limulus* cardiac activity has a nervous origin.

Stannius' ligatures.[2] In the amphibian heart the sinus, auricles, and ventricle beat successively one after the other. If a tight ligature is placed between the sinus and the auricles (Stannius' first ligature), the sinus continues to beat with its normal rhythm but the auricles and ventricle are stopped in diastolic relaxation. If a second ligature is placed between the auricles and the ventricle, rhythmic ventricular contractions reappear. Complete sections produce the same effects as the ligatures. The first explanation given to account for these facts was that Remak's ganglion, situated in the sinus, sends out rhythmic motor impulses that cause the whole heart to beat. A second nervous structure found in the auricle (Ludwig's ganglion) would emit inhibitory impulses, normally overcome by those of Remak's ganglion. Stannius' first ligature would release Ludwig's ganglion from the control of the sinus ganglion, and would therefore inhibit a third ganglion (Bidder's) situated in the ventricle. The latter would be released by Stannius' second ligature from the inhibitory impulses of Ludwig's ganglion and thus could start rhythmical ventricular contractions.

[1] CARLSON, A. J., *Am. J. Physiol.*, 12, 67, 1904.
[2] STANNIUS, H. F., *Arch. f. Anat. u. Physiol.*, p. 85, 1852.

The results are not always as typical as currently described in textbooks, and their explanation on the basis of the local ganglionic activity is no longer tenable. The results observed can be more satisfactorily explained by differences in automatism of the myocardium in different parts of the heart. Changes in the tension of the cardiac walls caused by the ligatures also influence the results.[1]

BIBLIOGRAPHY

BRONK, D. W., The Nervous Mechanism of Cardiovascular Control, *Harvey Lect.*, 29, 245, 1935.

DE CYON, E., "Les Nerfs du coeur," Librairie Félix Alcan, Paris, 1905.

HERING, H. E., "Die Karotissinusreflexe auf Herz und Gefässe," T. Steinkopf, Leipzig, 1927.

HEYMANS, C., "Introduction to the Regulation of the Blood Pressure and Heart Rate," Charles C Thomas, Springfield, Ill., 1950.

HEYMANS, C., J. J. BOUCKAERT, and P. REGNIERS, "Le Sinus carotidien et la zone homologue cardio-aortique," Gaston Doin, Paris, 1933.

KOCH, E., "Die reflectorische Selbsteuerung des Kreislaufes," T. Steinkopf, Leipzig, 1932.

McDOWALL, R. J. S., "The Control of the Circulation of the Blood," Longmans, New York, 1938.

[1] POSTMAN, N., *Experientia*, 7, 67, 1951.

CHAPTER 18

The Nutrition and Work of the Heart

THE FUNCTIONAL INTEGRITY of the heart depends in the first place on the nutrition of its fibers, which in mammals is assured by the blood flowing through the coronary arteries. The coronary circulation has certain peculiarities, and its disturbances may have fatal consequences. For this reason a thorough knowledge of its physiology is of great importance to the physician.

CORONARY CIRCULATION

Anastomoses[1] between the different branches of each coronary artery and also between the territories of the right and left coronary arteries may be easily demonstrated by a variety of anatomical procedures. From a functional point of view, however, the coronary arteries must be considered as terminal vessels. The anastomotic channels cannot compensate adequately for the sudden occlusion of a branch of a certain importance. Sudden occlusion of a coronary artery or of one of its branches causes death of the tissues it nourished, and an infarct is formed. If the subject does not die, scar tissue replaces the dead fibers in the area of necrosis. If occlusion occurs gradually, the anastomoses develop considerably and no infarction occurs when eventually the artery is completely occluded.

The *veins of the heart* run beside the arteries. The majority of the veins run toward, and end in, the coronary sinus, which opens into the lower part of the posterior wall of the right auricle. At this opening Thebesius' valve is found.

[1] PRINZMETAL, M., B. SIMKIN, H. C. BERGMAN, and H. C. KRUGER, *Am. Heart J.*, **33**, 420, 1947. ZOLL, P. M., S. WESSLER, and M. J. SCHLESINGER, *Circulation*, **4**, 797, 1951.

Embedded in the myocardium there are also small short vessels lined by endothelium, which open directly into the cavities of the heart. They are called veins of Thebesius. According to some investigators they communicate with the capillaries of the heart. They are of no value as an auxiliary circulatory system in cases of sudden coronary occlusion, a fact that has been shown by experiments on dogs and in cases of coronary occlusion in man.

Dynamics of the coronary circulation. The circulation in the coronary arteries has certain peculiar features owing to the origin of the arteries in the root of the aorta and their distribution in the walls of the heart, where they are subject to rhythmic compression and decompression at each systole and diastole. It seems definitely established that the velocity of flow decreases progressively, until it almost ceases during systole, and then rapidly increases from the beginning of diastole, to reach its peak in mid-diastole. The myocardium receives during systole approximately three-quarters of the volume of blood it receives during an equivalent period of the diastole, but as the diastole lasts longer than the systole, the proportion of blood that the myocardium receives during diastole is considerably greater (Wiggers).

The pressure in the main branches of the coronary arteries closely follows the changes in the aorta.[1] Figure 82 shows the pressure in the aorta and the arterial pulse in the ramus descendens anterior, simultaneously recorded in

[1] WIGGERS, C. J., and F. S. COTTON, *Am. J. Physiol.*, **106**, 9 and 597, 1939.

a dog. There is perfect coincidence between the curves.

The coronary circulation depends essentially on the difference in pressure between the aorta and the right auricle. It is influenced by the contraction and relaxation of the myocardium and is under the control of the autonomic nervous system, which governs the caliber of the vessels. The innervation of the coronary arteries is also peculiar inasmuch as sympathetic impulses produce dilatation in them, whereas in most other territories they produce vasoconstriction. Constriction of the coronary vessels is provoked by parasympathetic impulses, which travel along the fibers of the vagi nerves. Drugs produce effects according to their sympathicomimetic or parasympathicomimetic activity. For instance, acetylcholine (a parasympathicomimetic drug) produces constriction of the coronary arteries, and adrenaline (a sympathicomimetic drug), dilatation. Muscarine and pilocarpine have the same effect as acetylcholine; atropine, nitrites, and digitalis dilate the coronary vessels.

Coronary blood flow. The coronary blood flow can be measured experimentally by introducing a cannula directly into the coronary sinus and collecting the blood which flows out. This procedure has been used as a standard for calibrating and testing the accuracy of indirect methods that can be applied in man. For example, in the nitrous oxide method the subject breathes a certain amount of nitrous oxide which passes through the lungs into the arterial blood; an intracardiac catheter is introduced through a vein and then through the right auricle into the coronary sinus, and the concentration of nitrous oxide in arterial and in coronary venous blood is determined. The partition coefficient of nitrous oxide in heart tissue is first established by *in vitro* experiments, and with these data it is possible to calculate the blood flow through the coronary system. Simultaneous determinations made with the direct and the indirect (nitrous oxide) methods in dogs have given satisfactory results.[1] The nitrous oxide method has been applied in normal human subjects, although the procedure is not an easy one. The coronary blood flow has been found to be 65 cc./100 gm. of heart tissue per minute. This is 7 to 10 per cent of the total cardiac output per minute.[1]

[1] Eckenhoff, J. E., J. H. Hafkenschiel, M. H. Harmel, W. T. Goodale, M. Lubin, R. J. Bing, and S. S. Kety, *Am. J. Physiol.*, **152,** 356, 1948; Gregg, D. E., F. H. Longino, P. A. Green, and L. J. Czerwonka, *Circulation*, **3,** 89, 1951.

Coronary occlusion. The occlusion of a main branch of a coronary artery produces a serious disturbance in the functioning of the heart and is frequently fatal. Coronary occlusion is becoming an increasingly frequent cause of sickness and death and therefore occupies an ever-larger place in human pathology. It produces the dramatic syndrome known as angina pectoris (see page 161).

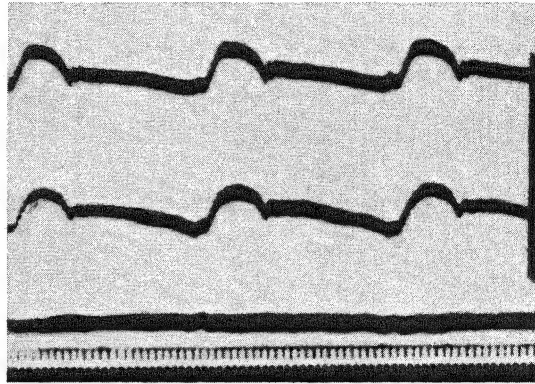



Fig. 82. Coronary pulse and aortic pressure in the dog. Upper tracing: sphygmographic record of the anterior descending branch of the left coronary artery. Lower tracing: pressure pulse of the aorta (Wiggers' manometer). Note that both curves are identical and can be superimposed. Time in 0.02 sec. (*Braun Menéndez and Orías.*)

In an anesthetized dog, with the chest opened under artificial respiration, if one of the principal coronary arteries (right or left circumflex, or descendens anterior) is ligated, the most prominent disturbances observed affect (*a*) the rhythm of the heart; (*b*) its dynamics; (*c*) its action currents.

The changes in cardiac rhythm are conspicuous. The most frequent nonspecific disturbance in rhythm is the appearance of premature contractions, nearly always of ventricular origin. They begin 5 to 7 min. after the occlusion, at first occasionally, and then in more or less prolonged series. When extrasystoles appear, ventricular fibrillation is imminent and if it occurs the dog dies. If it does not occur, the premature beats tend to diminish in frequency and after 15 or 20 min. only occasional ones are observed which soon disappear completely. Extrasystoles may arise in ectopic foci situated in the ischemic area,

[1] Bing, R. J., *et al.*, *Am. Heart J.*, **38,** 1, 1949.

or they may be due to local disturbances in conductivity which make a reentrance of the normal impulse possible.

Ventricular fibrillation is a frequent result of experimental coronary occlusion; it has occurred in 75 per cent of the experiments in dogs in which the anterior descending branch was ligated. Sometimes it suddenly puts an end to the activity of a heart that was beating vigorously and maintaining good circulatory conditions in spite of the infarct. In other experiments fibrillation is the final stage of a hypodynamic heart that is beating very feebly. It is always preceded by premature contractions.

Ventricular fibrillation following coronary occlusion is probably due to changed conditions of excitability and conductivity affecting the ischemic area. Furthermore, the cardiac fibers in the infarcted area are not all damaged to the same extent; therefore there are also differences in excitability and conductivity within the necrosed area. Ischemia alone, if it affects the whole heart uniformly, does not cause fibrillation, as is seen in a heart separated from the body and left without any circulation through its vessels; it simply ceases to beat and remains in diastolic relaxation.

Hemodynamic disturbances following experimental occlusion of a main artery or branch of the coronary arteries are not so evident on direct inspection as the disturbances in rhythm, but they have a marked effect on the circulation. They can be observed by recording with optical manometers the intraventricular and aortic pressures; their severity can thus be measured and the compensatory reactions, when they occur, can also be followed.[1]

If the branch occluded is important because of the amount of heart tissue it nourishes, in the majority of cases the heartbeat commences to weaken progressively, immediately after the artery has been tied. Intraventricular and aortic pressures diminish, ventricular systole is shortened, especially the ejection phase, and the

Fig. 83. Coronary occlusion and adequate compensatory reaction. Aortic pressure, left intraventricular pressure, ECG (lead II). *A*, before ligation of the anterior descending branch of the left coronary artery in a dog; *B* to *F*, at different intervals after ligation. Discussion in text (*Orías, O., Am. J. Physiol., vol. 100, p. 629, 1932.*)

[1] Orías, O., *Am. J. Physiol.*, **100**, 629, 1932; Braun Menéndez, E., and O. Orías, *Rev. Soc. argent. de biol.*, **10**, 14, 1934; Maldonado-Allende, I., and O. Orías, *Rev. Soc. argent. de biol.*, **12**, 279, 1936.

initial ventricular tension increases (Fig. 83, *B* and *C*). The subsequent events vary in different cases. In some of them, after a few minutes, the aortic and intraventricular pressures gradually recover the normal level and the ejection phase lengthens, so that the duration of systole is restored and the contour of the pressure curves appears the same, or almost the same, as in the beats preceding occlusion (Fig. 83, *D*, *E*, and *F*). The increase in initial ventricular tension determines the more vigorous contraction of the myocardial fibers not submitted to ischemia (Starling's law of the heart) (Fig. 84). A compensated hypodynamic reaction has occurred. If compensation is complete, as far as the dynamics of the heart is concerned, and ventricular fibrillation has not occurred, the heart continues to function efficiently for a long time, and the animal survives.

If the area subjected to ischemia is too large, or if the other parts of the myocardium are already damaged, hypodynamia becomes progressively more marked, in spite of the increased initial ventricular tension (Fig. 85). On direct examination, the ventricles are seen to be dilated, and there is pronounced stasis in the auricles and large veins. After a short time ventricular fibrillation puts an end to the activity of a heart already incapable of maintaining the circulation. This sequence of events is observed in cases of uncompensated progressive hypodynamia.

In a small number of cases overcompensation may occur. Intraventricular and aortic pressures rise above the level previous to the occlusion, and the duration of systole is prolonged; there is also an increase in initial ventricular tension.

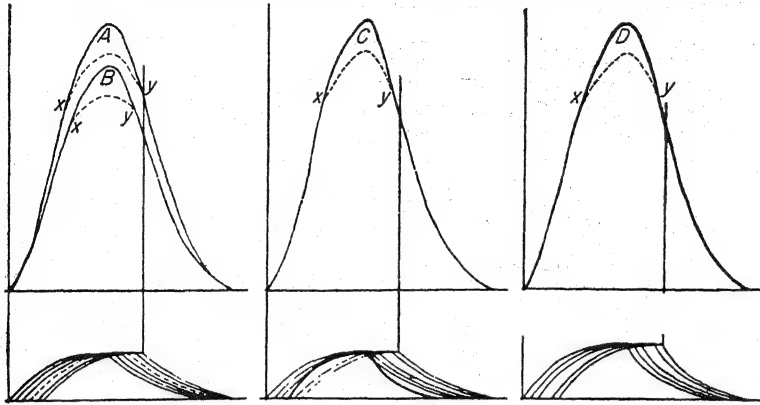



FIG. 84. Diagram to explain the response to sudden coronary occlusion. The upper curves represent the algebraic sum of the lower curves. Each one of the latter represents the contraction of a myocardial unit. Only six have been drawn. The curve *A* is the sum of the six curves below it. *B* is the sum of only five; the unit represented by the small curve in dotted line is supposed not to have responded (hypodynamic contraction). *C* is the sum of the six curves below it; it is of shorter duration because the unit represented by the thick-line curve is supposed to have been modified by ischemia (abbreviated systole). *D* is approximately the same as *A* but is the sum of only five unit curves, which are higher than those under *A*, *B*, and *C* (adequate compensatory reaction). (*Orías, 1932.*)

These experiments show that a myocardial infarct cannot be diagnosed by the changes in arterial blood pressure. A normal, or even a high, blood pressure is not incompatible with the existence of necrotic foci in the myocardium

Disturbances in the action currents of the heart caused by coronary occlusion are revealed principally by changes in the ST segment of the ECG, which after occlusion is usually above or below the isoelectric level, and in the T wave, which may appear deformed or reversed. These changes are typical, within certain limits, for the occlusion of each one of the arterial branches, but they are not apparent in all the leads. The location of the infarcted area conditions the type of change in the ECG. The extrinsic innervation of the heart has no influence on the variations in the ECG. Precordial leads are especially useful for discovering the electrocardiographic changes caused by coronary occlusion. These changes appear within the first two minutes after the vessel has been tied, and they disappear if the ligature is removed a few minutes later (Fig. 86). Venous occlusion alone does not produce them.

Changes in the ECG are the most important signs for the diagnosis of myocardial infarction

and for following its course in man. In some cases these changes even permit the localization of the infarct. Figure 87 reproduces the ECG of a patient with coronary occlusion; in this case there was also auricular fibrillation.

have found that oxygen consumption of the human heart averages 7.8 cc./100 gm. of cardiac tissue per minute.

The oxygen content of coronary venous blood in man and in the dog is considerably less than

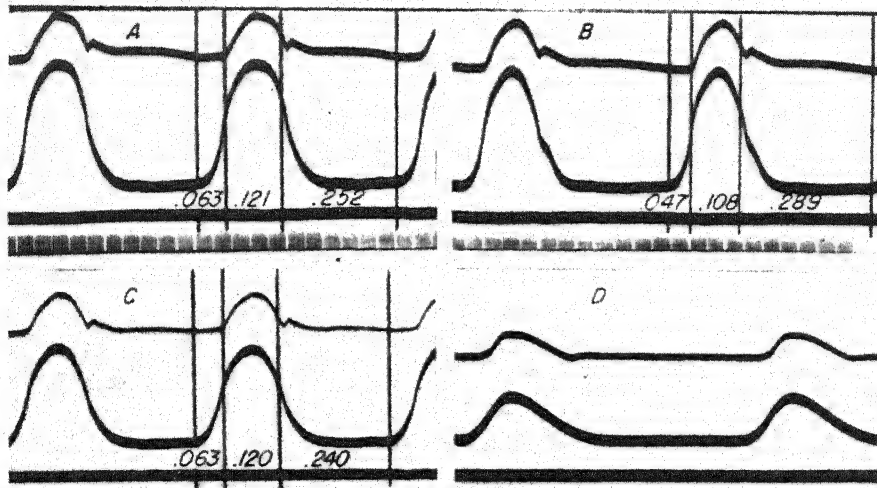



FIG. 85. Coronary occlusion. Hypodynamic uncompensated response. *A*, aortic and left intraventricular pressure before ligation of the anterior descending branch in a dog. *B*, *C*, and *D* were taken after occlusion of the artery; note the progressive decrease in the strength of the contraction. (*Orías, 1932.*)

that of the mixed blood in the right ventricle. A low oxygen concentration in the sample of blood collected, which can be seen by the dark color of the blood, is a sign that the catheter is rightly placed in the coronary sinus. Bing *et al.*[1] found 11.6 to 15.6 volumes of oxygen per 100 cc. in the blood of the right ventricle, and only 3.9 to 6.9 volumes per 100 cc. in coronary sinus blood of the same subjects. Cardiac tissues, therefore, remove relatively more oxygen (12 volumes per 100 cc., according to Bing *et al.*) than any other tissue in the body. Oxygen consumption per unit weight of cardiac tissue is normal in patients with arterial hypertension.

These determinations have been made in resting subjects. In experiments made on intact animals or with isolated hearts it has been proved that oxygen consumption increases with the work of the heart. Adrenaline also stimulates considerably cardiac oxygen consumption.

THE METABOLISM OF THE HEART

Energy for the contraction of the heart muscle is ultimately derived from oxidation processes which entail the consumption of oxygen and production of CO_2. The total metabolism of the heart can be measured by determining the oxygen consumption or the production of CO_2, but it is also important to know the nature of the oxidized substances and the mechanism of their oxidation. Energy can also be obtained from enzymatic chemical reactions when there is no oxygen available. Unfortunately little is known about these metabolic processes. A thorough knowledge of them would undoubtedly be of great importance in medical practice. Only a brief summary of the better known aspects of cardiac metabolism will be given here.

Oxygen consumption. Oxygen consumption can be measured by determining the difference in the oxygen percentage of arterial and venous blood in the coronary system, and the amount of blood that passes through the heart in unit time. This method has been applied in man, using an intracardiac catheter, the tip of which is introduced into the coronary sinus, but this procedure is by no means easy. Bing and his associates[1]

Heart-lung preparation.[2] Important data on cardiac metabolism can be obtained from the heart-lung preparation. The oxygen consumption of the preparation is the sum of the oxygen

[1] BING, R. J., *et al.*, *Am. Heart J.*, **38**, 1, 1949.

[1] *Ibid.*

[2] STARLING, E. H., *et al.*, *J. Physiol.*, **40**, 285, 1910; **44**, 206, 1912; **48**, 357, 1914.

consumption of the heart and of the lungs. The oxygen consumed by the lungs can be determined by running a control; by deducting it from the total oxygen consumption, the figure corresponding to the heart will be obtained. The heart-lung preparation has also been used to obtain information concerning the substances used as fuel by the heart.

The pulmonary circuit can also be replaced by an artificial oxygenator[1] circuit, in order to avoid the complicating influence of the metabolic processes of the lungs. The heart-lung and heart-oxygenator preparations have the advantage over simple perfusion of the isolated heart of making the heart perform work in conditions more similar to normal.

The fuel used by the heart. Only incomplete data have so far been obtained on the substances used as fuel by the heart. The usual procedure consists in measuring the amount removed from the blood circulating in a heart-lung preparation, or better still in a heart-oxygenator preparation, of the substance studied. Perfusion of the isolated heart with blood, or with a fluid in which the concentration of the substance is known, may also be used in these studies. The determination of the respiratory quotient (ratio of CO_2 produced to O_2 consumed) of the isolated heart, or of the heart-lung preparation, also gives information on the nature of the substances burned.

A well-established fact in cardiac metabolism is the utilization of large amounts of lactic acid, which seems to be the glucidic substance used preferentially to all others by the heart. In ordinary conditions, when there is a good supply of oxygen there is little consumption of glucose, but there is a considerable consumption of lactate. The heart of a dog in a heart-oxygenator preparation consumes 70 mg. of glucose and 200 mg. of lactate per 100 gm. per hr. When the work of the heart increases, the consumption of both these substances also increases. A decrease in the supply of either increases the consumption of the other. The lactic acid used by the heart has its origin in the process of glycolysis that normally takes place in the blood; this process is greatly activated when the blood passes through the lung. Lactic acid produced in the contraction of skeletal muscle and diffused into the blood is also a source of energy for the myocardium.

[1] Lovatt-Evans, C., L. Grande, and F. Y. Hsu, *Quart. J. Exper. Physiol.*, **82**, 41, 1934.

A
R
P
T
B
R
P
T
C
P R
T
D
P
R
T
E
R
P
T

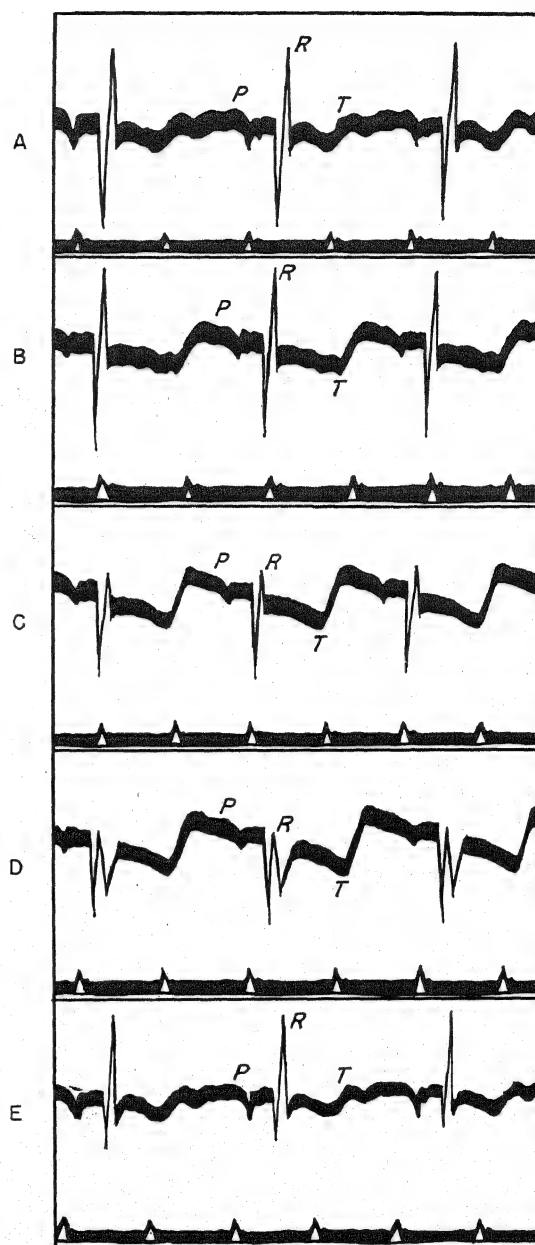



Fig. 86. ECG in experimental coronary occlusion (lead IV, chest-back), taken before and after occlusion of the anterior descending branch in a dog. *A*, before the occlusion; *B*, 2 min. after the occlusion; *C*, 3 min. after; *D*, 5 min. after; *E*, 10 min. after removing the ligature. Note the progressive disturbance in the ECG, and complete recovery after the circulation has been restored. Time in 0.2 sec. (*García del Río and Orías, Rev. Soc. argent de biol., vol. 10, p. 148, 1934.*)

If the glucose and lactate in the blood are exhausted, the contraction of the heart is considerably weakened and the glycogen content of the myocardium diminishes. If either glucose or lactate is again added to the blood, the heartbeats recover their vigor, lactate being more efficient than glucose. On the other hand myocardial glycogen is restored by adding glucose to the blood, but not by adding lactate. Monoiodoacetic acid, which rapidly paralyzes skeletal muscle, does not modify cardiac activity if there is a sufficient oxygen supply.

When there is an insufficient oxygen supply the nutritive processes of the heart undergo a marked change: the myocardium no longer utilizes lactate, but discharges it into the blood, and monoiodoacetic acid rapidly stops the heartbeat. The metabolism of the heart muscle in anoxia thus resembles the metabolism of skeletal muscle.

The following interpretation seems to cover the facts so far known: energy for the contraction of the heart fibers is provided directly or indirectly by the disintegration of phosphocreatine; energy for the resynthesis of phosphocreatine is provided by the oxidation of lactate. In conditions of anoxia, energy for the resynthesis of phosphocreatine is provided by the disintegration of glycogen into lactic acid (as in skeletal muscle).[1]

The heart can utilize pyruvic acid,[2] which is chemically a close relative of lactic acid. It is also supposed to oxidize fats, as the respiratory quotient of the heart deprived of glucose and lactate is 0.7 (RQ of fat oxidation). After a heart-lung preparation has been functioning for some time, the fat content of the heart diminishes. When β-hydroxybutyric acid (one of the intermediate substances in the metabolism of fat) is added to the heart-lung preparation, it is removed from the blood.

There is not much information on the protein metabolism of the heart; but the myocardium does not seem to make direct use of amino acids. Glycine added to the blood circulating in a heart-lung preparation increases the formation of creatine.

Vitamin B_1 (thiamine) plays an important part in cardiac metabolism. Serious and even fatal disturbances in the heart have been ob-

FIG. 87. ECG of a case of coronary occlusion in man. T starts at S in lead I and from the descending limb of R in leads II and III. This type of record is interpreted as corresponding to ischemia of the base and posterior wall of the ventricles (occlusion of the right coronary artery). Auricular fibrillation is indicated by the absence of P, the fine oscillations between the ventricular complexes, and the irregular intervals between these complexes. (*Courtesy of Dr. R. A. Brandán.*)

[1] LOVATT-EVANS, C., in E. H. Starling, "Principles of Human Physiology," 9th ed., J. & A. Churchill, London, 1945.

[2] BRAUN MENÉNDEZ, E., A. L. CHUTE, and R. A. GREGORY, *Quart. J. Exper. Physiol.*, **29**, 91, 1939.

served in the course of B_1 avitaminosis[1] (see Chap. 49).

The marked changes observed in the functions of the heart in both hyperthyroidism and hypothyroidism clearly show the importance of the internal secretion of the thyroid in normal cardiac metabolism. Adrenaline definitely increases the oxygen consumption of the heart. Little is known about the influence of other internal secretions on cardiac metabolism.

THE WORK OF THE HEART

Energy set free in the course of cardiac metabolic processes is used in part to perform work, but the greater part of this energy is given off as heat. About three-quarters of the total energy set free is lost as heat.

The work of the heart consists in

1. Ejecting into the arteries a certain amount of blood against the resistance of the blood already in the aorta and pulmonary artery. In these arteries there is a certain pressure maintained by the elastic reaction of the partially distended arterial walls.
2. Imparting velocity to the blood ejected so that it will circulate.

The work performed by each ventricle at every beat can be expressed as follows:

$$W = QR + \frac{wv^2}{2g}$$

where W is work, Q is the amount of blood ejected at each systole in cubic centimeters, R is the mean arterial pressure in meters of water (pressure in centimeters of mercury multiplied by 13.6, which is the specific weight of Hg) at the level of the aorta or pulmonary artery, w is the weight of the blood ejected at each systole, v is the velocity of the blood stream in the aorta or pulmonary artery, and g is the gravity constant, *i.e.*, 9.81.

QR represents the work necessary to eject the blood against the arterial resistance, and $wv^2/2g$ the kinetic energy necessary to impart velocity to the column of blood. The result is expressed in gram-meters; one gram-meter is the work performed in raising 1 gm. a height of 1 m.

All data necessary for the calculation of the work performed by the heart can be obtained in man by methods that are in common use. Thus Q is given by the minute volume and the heart rate; R is determined by measuring the arterial blood pressure by any standard method if the resistance in the aorta has to be known, or by transcardiac catheterization in the case of resistance in the pulmonary artery. The kinetic component is calculated as follows: The velocity of the blood is directly proportional to the systolic discharge Q and inversely proportional to the area of the cross section of the aorta, or the pulmonary artery, and the duration of the ejection phase T. With these data the following formula for the kinetic energy has been deduced,[1] in which K is a constant and A the area of the cross section of the aorta (or pulmonary artery).

$$\text{Kinetic energy} = \frac{Q^3}{A^2 \times T^2} = K$$

In man the necessary data are obtained by intracardiac catheterization and angiocardiography.

In six normal, resting subjects the following figures, in gram-meters per beat, were obtained: potential energy for the right ventricle, from 8.42 to 19.70; for the left ventricle, 91.0 to 190.0; kinetic energy, for the right ventricle, 0.31 to 1.30; for the left ventricle, 0.46 to 3.90.

When the subject is resting, kinetic energy is only a small fraction of the total energy developed by the heart, but it increases in importance (up to 25 per cent of the total energy) during physical exercise.

Efficiency of the heart. The efficiency of a motor is the ratio between the energy converted into work and the total energy.

The efficiency of the heart can be calculated, approximately, not only in animals, but also in man.[2] The work of the heart is measured, and the total energy developed is calculated from the oxygen consumption of the heart, determined by the methods described in preceding paragraphs.

The energy cost is obtained by multiplying the oxygen consumption by its energy equivalent. If the RQ (respiratory quotient; see Chap. 40) for the heart is considered to be 0.82, the energy equivalent of oxygen would be 2,059 kg.-m. per liter. Oxygen consumption is usually calculated per 100 gm. of cardiac tissue, but the actual weight of the subject's left ventricle can be obtained from special anatomical tables; thus the total oxygen consumption for the ventricle can be calculated. The results are probably accurate only within ±15 per cent. The effi-

[1] Soldati, L. de, Los Trastornos Circulatorios de la Avitaminosis, thesis for M.D., Buenos Aires, 1940.

[1] Prec, O., L. N. Katz, L. Sennet, R. H. Rosenman, A. P. Fishman, and H. Hwang, *Am. J. Physiol.*, **159**, 483 1949.

[2] Bing, *et al.*, *loc. cit.*

ciency of the heart is given by the ratio

$$\frac{\text{Cardiac work}}{\text{Total energy cost}} \times 100$$

Bing and his associates,[1] applying these methods in normal men, found the efficiency to be 19.0 to 24.5 per cent, and in cases of congestive heart failure only 13 to 17 per cent.

THE OUTPUT OF THE HEART. THE MINUTE VOLUME

From many points of view it is important to know the output of the heart per minute.

Fick's principle.[2] The minute volume can be calculated by means of Fick's method; the oxygen intake by the lungs during 1 min. is measured, and the oxygen concentration in arterial blood and in the venous blood of the right ventricle is determined. The amount of blood in which the oxygen has been absorbed is easily calculated; this is the amount of blood passing through the lung in 1 min., *i.e.*, the minute output of the right ventricle. The left ventricle ejects the same amount as the right ventricle.

For example, a subject absorbs 248 cc. O_2 in 1 min. and the arterial blood has 45 cc. more oxygen per liter than the venous blood; therefore

$$\frac{1{,}000}{45} = \frac{x}{248}$$

$$x = \frac{248 \times 1{,}000}{45} = 5{,}511 \text{ cc.}$$

The right ventricle in this case has ejected 5,511 cc. in 1 min., and the left ventricle has ejected the same amount. Dividing this by the heart rate, the average systolic discharge is obtained.

The minute volume can be calculated in the same way by measuring the elimination of CO_2 and the arteriovenous difference in CO_2 concentration.

Oxygen or CO_2 concentration must be determined in the blood of the right auricle or ventricle, because this is blood entering the lung. Venous blood from different territories may differ considerably in its O_2 or CO_2 concentration, because of differences in the activity of the tissues. Arterial blood, on the other hand, may be taken from any artery, because it has the same composition in all the arteries.

[1] *Ibid.*

[2] Fick, A., *Sitzungsb. d. physik.-med. ges. sch. z. Würzburg*, p. 16, 1870.

Determination of the minute volume in man. Until rather recently it was customary to resort to indirect methods to determine the minute output of the heart in human subjects. Since the systematic researches of Cournand and his associates,[1] however, the *direct application of Fick's principle* is becoming more and more the procedure of choice. Venous blood is drawn from either the right auricle or the right ventricle by means of a catheter introduced into the heart chambers through one of the veins of the arm. Arterial blood is obtained by puncture of the radial or any other superficial artery. The subject's oxygen consumption is determined by a standard method (see Chap. 40). Samples of blood must be drawn under paraffin while the oxygen consumption is being determined.

There are other indirect methods for measuring the minute volume which can be applied in man. In the acetylene[2] and nitrous oxide[3] methods the subject breathes the gas. The minute volume is calculated from the amount of gas taken up by the body and its concentration in the blood. In other methods the minute volume is calculated from the distribution in the blood of special dyes.[4] All these methods require rather elaborate techniques and are not exempt from errors.

Intracardiac catheterization. Forssmann[5] was the first to apply venous catheterization of the heart in man; after studying the technique in the cadaver, he first performed it on himself. Other investigators occasionally made use of this method,[6] but the credit for systematic use of venous catheterization of the heart is due to Cournand and his collaborators in the United States, and to McMichael and his collaborators in England.[7]

The skin over a vein in the arm is anesthetized and

[1] Cournand, A., and H. A. Ranges, *Proc. Soc. Exper. Biol. & Med.*, **46**, 462, 1941.

[2] Krogh, A., and J. Linhard, *Skandinav. Arch. f. Physiol.*, **27**, 100 and 227, 1912.

[3] Marshall, E. K., and A. Grollman, *Am. J. Physiol.*, **86**, 117, 1928; **88**, 432, 1929.

[4] Hamilton, W. F., *et al.*, *Am. J. Physiol.*, **153**, 309, 1948.

[5] Forssmann, W., *Klin. Wchnschr.*, **8**, 2085, 1929.

[6] Padilla, T., P. Cossio, and I. Berconsky, *Semana méd.*, 2, 79, 1932.

[7] McMichael, J., and E. P. Sharpey-Shäfer, *Brit. Heart J.*, **6**, 33, 1944.

cleaned thoroughly, and a flexible ureteral catheter (gauge 12) is introduced into the vein. The catheter is slightly bent at 6 to 8 cm. from the end, as this facilitates its entrance into the auricle. If the catheter is opaque to x-rays, its progress can be followed and its tip accurately guided. The injection of a small quantity of sterilized sodium iodide solution, which is also opaque to x-rays, permits the visual control of the position of the catheter, which can be passed to the ventricle if it is desirable to do so. The catheter is flushed with sterilized saline solution and is then connected with a flask containing heparin in saline. A manometric tube graduated in centimeters gives the pressure in the system. Sterile saline is passed continuously through the catheter at the rate of 50 cc. per hr., except when samples of blood are drawn. Not only is the position of the end of the catheter controlled by x-rays; the nature of the blood drawn and the type of the pressure waves registered by a suitable manometer connected to the catheter also give valuable information for this purpose. The cardiac rate and rhythm should be followed in the ECG throughout the procedure. The repeated appearance of extrasystoles is a sign of undesirable irritation of the internal surface of the heart. The nature of an extrasystole will also give information on the position of the tip of the catheter. An already wide experience has proved that this method is usually well tolerated by the subjects submitted to it, but it must not be forgotten that there is a certain risk of thrombosis and ventricular fibrillation.

This method was originally used for intracardiac exploration, but it has since been applied to the collection of blood from deeply situated vessels, *e.g.*, the coronary sinus, the hepatic or renal veins (the catheter is passed through the auricle into the inferior vena cava), and the pulmonary artery or its branches (through the right auricle and ventricle).

The left cavities of the heart have been explored by the introduction of the catheter into the ulnar artery. This is not as easy as venous catheterization, owing to the high pressure in the arteries and the obstacle represented by the aortic valves.[1]

Normal minute volume in man. The average minute volume of normal subjects in basal conditions is 5.4 liters/min., when it is determined by the application of Fick's principle. Different workers have obtained results that are generally in accordance with this average. The minute volume per square meter of body surface is 3.1 liters; this is known as the *cardiac index*.

When the subject stands up, the minute volume diminishes (from 5.3 liters lying down to 4 liters standing up, according to McMichael and Sharpey-Schäfer[1]). Exercise increases the minute volume in proportion to the amount of exercise, up to 15 and 20 liters per min.

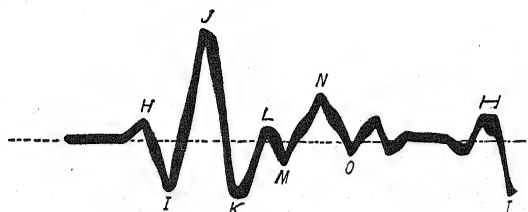



FIG. 88. Normal ballistocardiogram. *H*, apex beat; *I*, ballistic recoil during the ejection phase; *J*, impact of blood on the large vessels; *K*, impact of blood on peripheral resistance. The significance of the other waves is not yet clear. (*Sharr, I., and H. A. Schroeder, J. Clin. Investigation, vol. 19, p. 437, 1940.*)

The minute volume is the result of the systolic discharge multiplied by the heart rate. Factors that modify the minute volume will be considered in detail when discussing the importance of the minute volume in maintaining the arterial blood pressure.

Ballistocardiography. If a person lies on a special platform adequately mounted on springs, the platform oscillates synchronously with the heartbeats. The movements are presumably due to the ballistic recoil of the heart during the ejection phase and to the impact of the blood against the aortic arch. The movements of the platform can be amplified and recorded (Fig. 88). The records are called "ballistocardiograms," and by applying certain mathematical formulas it is possible, according to the proponents, to calculate from them the output of the heart.[2] The aspect of the waves recorded in the ballistocardiogram is studied for diagnostic purposes, because although there are no typical curves for the different disturbances in cardiac function, certain alterations in the strength and form of systolic ejection may provoke changes in these waves. A diagnostic conclusion cannot, however, be supported by the ballistocardiogram alone.

BIBLIOGRAPHY

Coronary Circulation

BATTRO, A., "Síndrome de Oclusión Coronaria," El Ateneo, Buenos Aires, 1930.

[1] ZIMMERMAN, H. A., R. W. SCOTT, and N. O. BECKER, *Circulation*, 1, 357, 1950.

[1] MCMICHAEL and SHARPEY-SCHÄFER, *loc. cit.*

[2] STARR, I., A. J. RAWSON, H. A. SCHROEDER, and N. R. JOSEPH, *Am. J. Physiol.*, 127, 1, 1939.

BING, R. J., The Coronary Circulation in Health and Disease, as Studied by Coronary Sinus Catheterization, *Bull. New York Acad. Med.*, **27**, 407, 1951.

CONDORELLI, L., "Die Ernährung des Herzens und die Folgen seine Störungen," T. Steinkopf, Leipzig, 1932.

GREGG, D. E., *Physiol. Rev.*, **26**, 28, 1946.

GROSS, L., "The Blood Supply of the Heart," Hoeber, New York, 1921.

HOCHREIN, M., "Der Koronarkreislauf," Springer, Berlin, 1932.

HOCHREIN, M., "Der Myokardinfarkt," 3d ed., T. Steinkopf, Leipzig, 1945.

LEVY, R. L., "Diseases of the Coronary Arteries and Cardiac Pain," Macmillan, New York, 1936.

SPALTEHOLZ, W., "Die Arterien der Herzwand," S. Hirzel, Leipzig, 1924.

WIGGERS C. J., in R. L. Levy, "Diseases of the Coronary Arteries and Cardiac Pain," Macmillan, New York, 1936.

WIGGERS, C. J., The Functional Importance of Coronary Collaterals, *Circulation*, **5**, 609, 1952.

The Metabolism of the Heart

BRAUN MENÉNDEZ, E., "Patología y Clínica de Enfermedades del Aparato Circulatorio," Conferencias, Círculo Médico de Córdoba, 1943, p. 3.

CLARK, A. J., M. G. EGGLETON, R. EGGLETON, G. GADDIE, and C. P. STEWART, "The Metabolism of Frog's Heart," Oliver & Boyd, Edinburgh and London, 1938.

CRUIKSHANK, E. W. H., *Physiol. Rev.*, **16**, 597, 1936.

LOVATT-EVANS, C., in W. H. Newton, "Recent Advances in Physiology," 6th ed., J. & A. Churchill, London, 1939, Chap. 4.

The Work of the Heart

WIGGERS, C. J., and L. N. KATZ, *Am. J. Physiol.*, **85**, 229, 1928.

Minute Volume

BERCONSKY, I., "El Volumen Circulatorio por Minuto en el Estado Normal y Patológico," El Ateneo, Buenos Aires, 1930.

CHARLIER, R., *Acta cardiol.*, **1**, 136, 1946.

CHRISTIANSEN, E. H., *Ergebn. d. Physiol.*, **39**, 348, 1937.

GROLLMAN, A., "Cardiac Output of Man in Health and Disease," Charles C Thomas, Springfield, Ill,, 1932.

LEQUIME, J., "Le Débit Cardiaque," Masson et Cie, Paris, 1940.

SEELY, R. D., W. E. NERLICH, and D. E. GREGG, A Comparison of the Cardiac Output Determined by the Fick Procedure and a Direct Method Using the Rotameter, *Circulation*, **1**, 1261, 1950.

Ballistocardiography

DOCK, W., H. MANDELBAUM and R. A. MANDELBAUM, "Ballistocardiography," Mosby, St. Louis, 1953.

DOCK, W., H. MANDELBAUM and R. A. MANDELBAUM, Ballistocardiography in Medical Practice, *J. A. M. A.*, **146**, 1248, 1951.

CHAPTER 19

Circulation in the Arteries

THE ARTERIES ARE tubes with elastic and muscular walls; they carry the blood to all the organs and tissues of the body. Their structure changes as their distance from the heart increases. The principal change concerns the relative importance of the elastic and muscular coats. Whereas the aorta is almost entirely made up of the elastic tissue, the small arteries and arterioles have a thick muscular layer. These structural differences assure a greater functional efficiency of the arteries in the several parts of the circulatory system.

The walls of the large arteries have great strength and elasticity, and therefore they can resist the high pressure to which they are submitted. Moreover during systole they store part of the systolic energy by converting it into tension, which is reconverted into kinetic energy during diastole. In small arteries, contraction and relaxation of the muscular coat are under the control of nervous and humoral factors, which regulate the inflow of blood to the different tissues according to their physiologic needs.

The aorta, because of its elastic walls, acts as a compression chamber; it absorbs and damps the sudden systolic impulse and sends blood to the small arteries at a more uniform pressure, speed, and rate.

There is still some debate as to whether the contraction of the arterial muscle layer contributes to propel the blood. Some authorities[1] believe this is so, but undoubtedly the main force that makes the blood circulate is the contraction of the heart. The contribution of the arterial muscles, if it exists, must be of very little importance. On the other hand the arteriolar muscles are an important factor in the maintenance and regulation of arterial blood pressure and in the regulation of the blood flow through local territories.

[1] HÜRTHLE, K., *Pflüger's Arch. f. d. ges. Physiol.*, **236**, 385, 1935.

Angiocardiography.[1] The arteries and other blood vessels and the cavities of the heart can be made opaque to x-rays by rapidly injecting them with a concentrated solution of a substance which is opaque to x-rays and is well tolerated by the organism, *e.g.*, a compound of iodine, such as 70 per cent diodrast. Several x-ray films (six or more) are taken at short intervals, which give accurate information on the diameter of the blood vessel, its permeability, etc. The aorta can be visualized by injecting the contrast substance into one of the carotid arteries, directing the needle toward the heart (retrograde aortography).[2]

THE VELOCITY OF BLOOD FLOW IN THE ARTERIES

The velocity of the blood flow varies inversely with the cross-sectional area of the vascular bed in which it circulates. This area increases progressively as the arteries ramify and the distance from the heart increases. The cross-sectional area of each branch is smaller, but the sum of the cross-sectional areas of the branches is always greater than that of the parent vessel. The velocity of the blood flow therefore diminishes progressively as the distance from the heart increases. The ratio of velocity to cross-sectional area is expressed as follows:

$$V = \frac{Q}{A} = \frac{Q}{\pi r^2}$$

where V is velocity, Q is the blood flow in unit time, and A is the cross-sectional area—or πr^2 if this is circular, r being the radius of the circle (half the diameter of the vessel).

[1] SUSMAN, M. L., and A. GRISHMAN, in Dock and Snapper's "Advances in Internal Medicine," Interscience Publishers, Inc., New York, 1947.

[2] PEREIRAS, R., and A. CASTELLANOS, *Radiology*, **53**, 859, 1949.

The velocity of the blood varies at different moments of the cardiac cycle and is greater in the center of the blood stream than in the periphery; therefore the mean velocity is the one generally considered. The activity of the heart causes a velocity pulse in the same way as it creates a pressure pulse. Velocity is at a maximum during the ejection phase of ventricular systole and at a minimum at the end of diastole, except in the blood vessels of the cardiac coronary system. The blood flow never stops completely but only slows down during diastole. The oscillations in velocity caused by the systolic ejection are damped as the distance from the heart increases and the diameter of the arteries decreases.

The differences in velocity between the center and the periphery of the blood flow are due to friction of the blood against the arterial walls and between the different concentric layers that form the blood stream itself. The greatest friction exists between the arterial wall and the most peripheral layer of blood; friction between the blood layers diminishes progressively from the periphery to the center of the stream. Therefore velocity is at a maximum in the center and diminishes progressively toward the periphery. This type of flow is observed in homogeneous fluids, and blood is not homogeneous, as it is made up of cells and plasma. The blood cells in circulating blood form columns in the center of the stream, and plasma circulates preferentially in the periphery. The greater viscosity of the blood cells diminishes the difference in velocity between the center and the periphery, which would be greater in a homogeneous fluid.

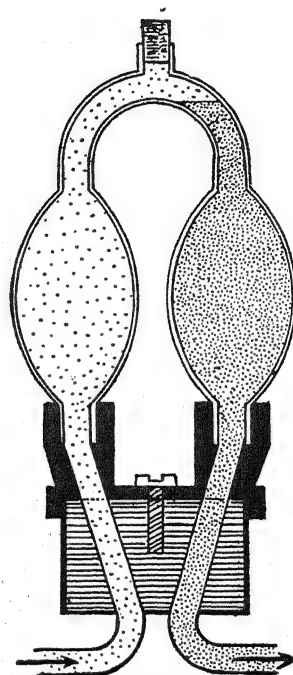



Fig. 89. Volkmann's hemodromometer.

Fig. 90. Ludwig's stromuhr.

Determination of the blood flow and its velocity. There are direct and indirect methods for determining the blood flow, *i.e.*, the amount of blood passing through a vessel in unit time, and its velocity. In the direct methods an instrument must be inserted into the blood vessel. Volkmann's hemodromometer was one of the first instruments used for the direct determination of the speed of blood flow and Ludwig's stromuhr for the direct measurement of flow and subsequent calculation of its velocity.

Volkmann's hemodromometer[1] (Fig. 89) consists of a U-shaped glass tube filled with saline solution. The time taken by the blood to fill the tube,

[1] Volkmann, A. W., "Die Hemodinamik," Leipzig, 1850.

displacing the saline solution, is measured with a chronometer; thus the speed of flow can be determined. The inertia of the saline in the tube introduces a certain error into this determination.

Ludwig's stromuhr[1] (Fig. 90) measures the amount of blood that passes through an artery in unit time. To calculate the speed of flow, the cross-sectional area of the artery must also be known.

Registration of the blood flow and its velocity (hemodromography). Indirect methods permit continuous registration of the speed of blood flow by means of instruments called hemodromographs. Oscillations in velocity can thus be continuously recorded and correlated with other phenomena occurring at the same time. If the instruments are calibrated, the blood flow can also be recorded.

Rein's stromuhr[2] registers the speed of flow by an indirect method, which does not require the insertion of an instrument into an artery. The artery is warmed at a certain place by means of diathermy (high-frequency electric waves), and the temperature of the blood is recorded by a thermocouple at a certain distance downstream from the spot warmed. The temperature of the blood will rise as the velocity of flow decreases; the instrument is calibrated so that temperature corresponds to speed of flow.

The mean rate of blood flow through a particular blood vessel can be registered by a recording rotameter[3] (Fig. 91). The cannulas *A* and *B* are inserted into the proximal and distal ends of the blood vessel. A metal float (*D*) rises more or less according to the amount and speed of blood flow. Its position provokes changes in an electromagnetic circuit (*C*), and these changes are registered by a recording galvanometer. The blood must be prevented from clotting by heparin or another anticoagulant.

Changes in the velocity of blood flow. The blood flows in the arteries at a greater speed during systole than during diastole. In the common carotid artery the speed reaches 500 mm. per sec. during systole and slows down to 300 mm. per sec. during diastole. In resting human subjects, with data obtained by intracardiac catheterization, angiocardiography, and other procedures, the mean velocity of flow in the aorta has been found to vary in different individuals between 21.3 and 87.4 cm./sec., and from 33.1 to 63.5 cm./sec. in the pulmonary artery.[1]

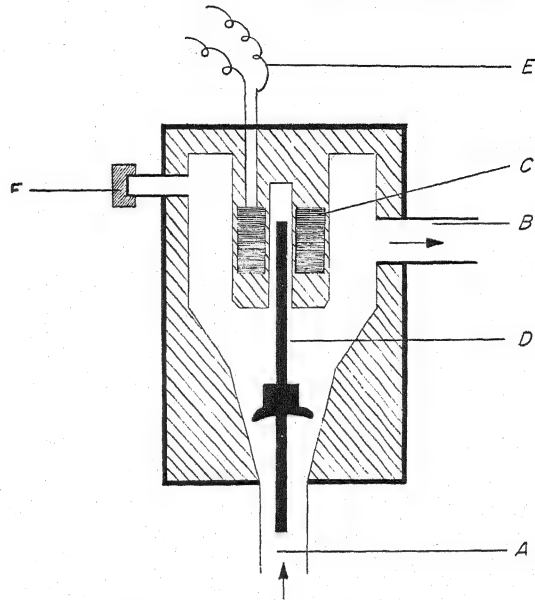



FIG. 91. Diagram of recording rotameter. *A*, connection with the proximal end of the blood vessel; *B*, connection with the distal end of the blood vessel; *C*, electromagnetic circuit; *D*, metal float, which on changing its position provokes changes in *C* transmitted by the cables *E* to a recording galvanometer; *F*, rubber stopper for cleaning and filling the instrument. (*After Shipley, R. E., and C. Wilson, Proc. Soc. Exper. Biol. & Med., vol. 78, p. 724, 1951.*)

Local changes in the diameter of the arteries can cause local variation in the velocity of the blood flow. Vasodilatation increases the velocity of the flow and vasoconstriction diminishes it.

THE BLOOD PRESSURE IN THE ARTERIES

The Reverend Stephen Hales, in 1732, gave the first demonstration that the blood in the arteries flows under pressure.[2] He described his experiment in the following words: "In December I caused a mare to be tied down alive on her back. Having laid open the left crural Artery

[1] DOGIEL, A. S., *Ber. Sächs Gesell.*, p. 199, 1867.

[2] REIN, H., *Ztschr. f. Biol.*, **87**, 394, 1928; CERLETTI, A., and E. ROTHLIN, *Helvet. physiol. et pharmacol. acta.*, **6**, 92, 1948.

[3] SHIPLEY, R. E., and C. WILSON, *Proc. Soc. Exper. Biol. & Med.*, **78**, 724, 1951.

[1] PREC, P., L. N. KATZ, L. SENNET, R. H. ROSENMAN, A. P. FISHMAN, and W. HWANG, *Am. J. Physiol.*, **159**, 483, 1949.

[2] Quoted by C. Lovatt-Evans in E. H. STARLING, "Principles of Human Physiology," 9th ed., J. & A. Churchill, London, 1945.

about three inches from her belly, I inserted into it a brass Pipe, whose bore was one sixth of an inch in diameter; and to that by means of another brass Pipe which was fitly adapted to it, I fixed a glass Tube, of nearly the same diameter, which was nine feet in length: Then untying the ligature on the Artery, the blood rose in the Tube eight feet three inches perpendicular above the level of the left Ventricle of the heart: But it did not attain to its full height at once; it rushed up about half way in an instant, and afterwards gradually at each Pulse, twelve, eight, six, four, two and sometimes one inch: When it was at its full height, it would rise and fall at and after each Pulse two, three, or four inches; and sometimes it would fall twelve or fourteen inches, and have there for a time the same Vibrations up and down at and after each Pulse, as it had when it was at its full height; to which it would rise again, after forty or fifty Pulses."

This experiment can be repeated in any animal. The height of the blood column balances the pressure in the artery.

Terminology. Arterial blood pressure is the force exerted by the blood on the walls of the arteries. The tension of the arterial wall changes with the pressure. The terms "arterial blood pressure" and "arterial tension" are therefore synonymous; the former is the one commonly used, generally in its abbreviated form of "blood pressure."

Poiseuille's law. Poiseuille's law[1] correlates the factors that determine the circulation of fluids in a system of tubes.

Poiseuille's law states that the volume of fluid (Q) passing through a system of tubes in unit time (minute or volume flow) is directly proportional to the difference between the pressure at the entrance to the system and that at its end ($P_1 - P_2$) and to the square of the cross-sectional area of the tube (q^2). It is inversely proportional to the length of the system (l), to the viscosity of the circulating fluid (η), and to a constant (8π); *i.e.*,

$$Q = \frac{(P_1 - P_2) \times q^2}{l \times \eta \times 8\pi}$$

Therefore the pressure difference will be

$$P_1 - P_2 = Q \frac{l \times \eta \times 8\pi}{q^2}$$

[1] Poiseuille, J. M., *Mem. Acad. Sc.*, **9**, 433, 1846.

where $(l \times \eta \times 8\pi)/q^2$ represents the factors conditioning resistance to the circulation of the fluid; it can be conveniently replaced by the letter R. Therefore

$$P_1 - P_2 = QR$$

In the circulatory system P_1 is the aortic pressure, and P_2 the pressure in the right auricle, which can be considered as zero. Therefore $P = QR$, *i.e.*, pressure in the aortic system is directly proportional to the minute volume ejected by the heart and to the peripheral resistance. Any change in these factors will cause a change in pressure in the same direction. If either of these factors is reduced to zero, the blood pressure will also fall to zero.

Factors that modify the minute volume. The minute volume of the heart is conditioned by the systolic discharge and the heart rate. When the systolic discharge remains constant, or diminishes only slightly, an increase in heart rate causes an increase in the minute volume and the arterial blood pressure rises. This occurs in moderate tachycardia, such as follows section of the vagus nerves, injection of atropine, or stimulation of the accelerators (stellate ganglion).

The systolic discharge is determined by several factors, some of which act in the opposite direction to others. The main factors are (*a*) the physiologic condition of the myocardium; (*b*) the venous return; (*c*) the heart rate.

The physiologic condition of the myocardium. This factor is of fundamental importance. If the heart muscle does not contract, or if its contraction is weak or uncoordinated, blood is not ejected into the arteries, or is ejected in insufficient amounts, and the arterial blood pressure falls; when there is no systolic discharge it falls to zero. The capacity of the myocardium to contract is affected by toxic (in some cases medicinal) substances; degenerative processes; anoxia (total, or limited to a restricted area, such as in the case of infarction by coronary occlusion); dystrophia such as that in B_1 avitaminosis; myocardial fatigue, etc.

Venous return. The amount of blood returning to the auricles by the veins is a primary factor conditioning an adequate systolic discharge. If blood does not return to the heart, however good the physiologic condition of the myocardium, it will not be able to pump into the arteries the blood which is not there. Several factors determine the venous return to the auricle. A

hemorrhage, for example, turns from its course a certain amount of blood. The venous return will then diminish; therefore the systolic discharge will be less, and the arterial blood pressure will fall. A similar effect results from general capillary dilatation, which is usually accompanied by increased permeability of the capillary membrane; more fluid passes from the blood plasma to the tissues, and the blood volume diminishes; moreover the velocity of the blood flow is diminished and the blood stagnates in the periphery. Transfusion of blood will have the opposite effect, but it must not be performed too rapidly; otherwise the capacity of adaptation of the myocardium is exceeded, and myogenic dilatation and acute cardiac insufficiency will follow. The flaccidity of the walls of the veins, and the low pressure of the blood circulating in them, make the veins easily collapsible, and even slight external compression will occlude them completely, thus interrupting the blood flow. If this occurs in the large thoracic veins the venous return will be hindered, and as the systolic discharge is diminished, the arterial blood pressure will fall. A decrease in thoracic pressure sufficient to dilate the veins and increase their capacity will have the same effect, because it causes venous stasis. Respiratory variations in intrathoracic pressure alternately produce both these effects on the great veins: compression during expiration, and decompression during inspiration. In the course of normal breathing the respiratory influence on venous circulation is not considerable; but a forced expiration with the glottis closed (Valsalva's experiment) or a forced inspiration in the same condition (Müller's experiment) will hinder the venous return and cause the arterial blood pressure to fall.

The heart rate. Since the heart rate conditions the duration of diastole, it also conditions the filling of the ventricle and the systolic discharge. In the course of intense tachycardia the increase in the number of systoles does not compensate the decrease in systolic discharge due to insufficient filling of the ventricle; therefore the minute volume diminishes and the blood pressure falls.

Factors that modify peripheral resistance. The peripheral resistance (R) is conditioned by several factors, correlated in Poiseuille's law by the expression $R = (l \times \eta \times 8\pi)/q^2$. The length of the system (l) is constant for each individual; the viscosity (η) is also fairly constant; and 8π is by definition a constant; the only variable factor is the cross-sectional area of the vessel (q). Therefore in a given individual, at a certain moment, the peripheral resistance will depend on the diameter of the blood vessels, which is under nervous and humoral control. Resistance is inversely proportional to the square of the cross-sectional area; therefore even a small decrease in the diameter of the arteries will cause a considerable increase in resistance, and therefore in arterial blood pressure.

The arterioles are in a permanent state of constriction, caused by impulses discharged from the vasomotor center. This tonic constriction is essential for a normal blood pressure; if it ceases (by inhibition or cocainization of the vasomotor center, or section of the spinal cord in the upper cervical region) the blood pressure falls below the level necessary to maintain an adequate circulation to all the tissues.

Permanent tonic constriction of the arterioles is governed by nervous and humoral mechanisms which cause changes in the diameter of the arteries, and thus modify the blood pressure. The effects of these factors will be considered in the next chapter.

Compensatory reactions. Arterial blood pressure is maintained at a fairly constant level by a complicated and sensitive mechanism of control. As soon as one of the factors mentioned modifies the cardiac output or peripheral resistance, compensatory reactions are provoked which tend to counteract the primary effect. For this reason marked changes in blood pressure are caused only by the vigorous action of these factors, and the effects, except in extreme cases, are usually transitory, even when the action persists. The different mechanisms which control blood pressure and their coordination will be considered in the next chapter.

QUANTITATIVE EVALUATION OF BLOOD PRESSURE

Hales's experiment, referred to in detail on page 177, was a quantitative determination of blood pressure. Hales's vertical tube acted as a manometer, but the clotting of the blood occurred too rapidly to allow a sufficiently prolonged observation.

Poiseuille's[1] application of the mercury manometer to the quantitative study of arterial blood

[1] POISEUILLE, J. M., "Recherches sur la force du coeur aortique," Paris, 1828.

pressure was an important advance. A mercury manometer is a U-shaped tube with long branches, partly filled with mercury. When both branches of the tube are submitted to the same pressure (*e.g.*, atmospheric pressure) the mercury is at the same level in both branches. If one of the branches is connected with an artery by means of a rubber tube and a cannula (the whole system is filled with anticoagulant fluid) the blood pressure will be transmitted to this branch of the manometer, and the level of mercury will descend in this branch and rise in the other. The difference in the level of the mercury in the branches measures the blood pressure in centimeters of mercury. This apparatus shows that the blood pressure fluctuates continuously without falling below a certain level.

FIG. 92. Ludwig's mercury manometer and kymograph.

Ludwig[1] made another important advance by placing a metal stem, provided with a float at one end, on the surface of the mercury in the open branch of the manometer. At the other end of the stem an indicator was attached so as to register on a kymograph the fluctuating level of the mercury column. This apparatus (Fig. 92), still in common use in physiological laboratories, was the first used to record circulatory phenomena, inaugurating thus the method of graphic registration, which has given such valuable results in medical and physiological research.

[1] LUDWIG, K., *Arch. f. Anat. u. Physiol.*, p. 242, 1847.

The mercury manometer does not reproduce faithfully the fluctuations in arterial blood pressure, because of its great inertia and low natural frequency. The fluctuations in pressure are therefore damped; the high values are less, and the low values greater, than the real ones. If the fluctuations are damped even more, by constricting the tube or rubber connection between the artery and the manometer as Marey proposed, only minute oscillations occur at each heartbeat, and a record of the mean arterial pressure is obtained. This is an important feature in circulation. The mean pressure would be the constant pressure which assures the same volume flow as the pulsating pressure. It could also be defined as the average of the infinite number of pressures existing during the interval considered. A damped mercury manometer gives an objective demonstration of the theoretical mean pressure and a reasonable approximation of its true value, without the need of a complicated mathematical calculation.

Elastic manometers are necessary to obtain a faithful record of the oscillations in arterial blood pressure; the most sensitive and accurate of them resort to optical registration. Wiggers' universal manometer and Hamilton's manometer give excellent results. They must be standardized with a mercury manometer to obtain the values of blood pressure.

The more accurate instruments also show that the pressure in the arteries is fluctuating continuously above a certain level, without ever falling to zero. Fluctuations usually occur above 60 to 70 mm. Hg. They are due to three main causes: cardiac activity, respiration, and vasomotor tone.

Cardiac fluctuations. The most important oscillations in blood pressure, because of their amplitude and medical significance, are those due to ventricular systole and diastole. The pressure rises from between 60 and 70 mm. Hg in diastole to between 110 and 130 mm. Hg at each systole. This sudden increase in pressure is due to the entrance into the aorta of the blood ejected by the ventricle, and is the cause of the arterial pulse. The exact form of this variation

in pressure at each cardiac cycle is shown in Figs. 93 and 94, which reproduce records obtained in human subjects with Hamilton's manometer, and in Fig. 106, a record obtained in the dog with Wiggers' manometer. The differences in the records obtained from different arteries will be discussed when considering the arterial pulse.

Respiratory fluctuations. Breathing causes variations in blood pressure. The maximum pressures observed during the three or four beats corresponding to inspiration are not the same as those in the three or four beats corresponding to expiration. The difference for the same moment of the cardiac cycle is usually of a few millimeters, seldom of 1 cm. (Fig. 93).

A double mechanism causes this respiratory variation in arterial blood pressure. On the one hand there is the functional correlation between the respiratory and cardiovascular centers in the medulla, and on the other the mechanical changes caused by the movements of the thorax. Functional correlation of the centers causes respiratory sinus arrhythmia (see page 145), with tachycardia during inspiration and bradycardia during expiration; therefore arterial blood pressure tends to rise in inspiration and fall in expiration. The mechanical effects of

FIG. 93. Record of femoral arterial blood pressure registered with Hamilton's manometer in man. Pressure scale on the left in mm. Hg. *R*, respiratory oscillations. (*Courtesy of Dr. A. Lanari.*)

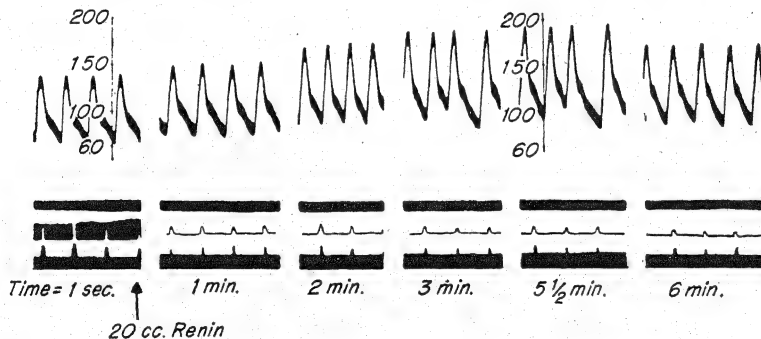



FIG. 94. Record of femoral arterial pressure in man, obtained with Hamilton's manometer. The arrow marks the injection of a dose of human renin. (*Battro, A., E. Braun Menéndez, A. Lanari, and L. F. Leloir, Rev. Soc. argent. de biol., vol. 16, p. 376, 1940.*)

respiratory movements produce the opposite results. During inspiration intrathoracic pressure falls, the thoracic veins are distended, and the venous return to the heart diminishes; the filling of the heart is reduced and the minute volume also diminishes; therefore the arterial blood pressure tends to fall. In the dog the effects of respiratory arrhythmia predominate (hypertension during inspiration); in man the mechanical effects are predominant (hypertension during expiration).

Vasomotor fluctuations. A third type of variation in blood pressure, much slower than the previous two, can be observed when a record is taken over a certain time. The line joining the peaks of the highest pulse wave in each respiratory cycle shows fluctuations, which have been attributed to variations in the tonic activity of the vasomotor centers.

The three types of fluctuation in arterial blood pressure can be seen in records taken with a mercury manometer, especially the cardiac and respiratory oscillations, but the true values of the pressure are not recorded.

In man the only fluctuations of practical interest are those due to ventricular contraction. It is usual in medical practice to measure the lowest pressure in diastole (diastolic pressure) and the highest in systole (systolic pressure). With the methods in common use other fluctuations can be detected only in special circumstances. The difference between systolic and diastolic pressures is known as the pulse pressure.

MEASUREMENT OF ARTERIAL BLOOD PRESSURE IN MAN

In man, the surgical methods used in animals to study the blood pressure are seldom employed. Direct determination of blood pressure can be made by puncturing an artery (brachial or femoral artery) through the skin, connecting the needle with a mercury manometer when the mean pressure is to be measured, or with a Hamilton manometer when a faithful record of the different oscillations is needed. In this last case the manometer and needle should be connected by a nondistensible but flexible tube (*e.g.*, thin lead tubing). The needle, tube, anticoagulant fluid, etc., should be carefully sterilized, but the whole operation is not difficult and entails no risk to the patient.

In current medical practice blood pressure is measured by methods that cause no inconvenience to the patients. They consist in exerting pressure through the skin and tissues on the artery in which the blood pressure is to be measured, by means of a pneumatic bag, applied to the limb by an inextensible cuff or bandage. An inflating bulb or pump serves to give the desired pressure, which is measured by a manometer. A valve permits gradual escape of the air in the bag, thus lowering the pressure. Mercury manometers are more accurate than aneroid manometers, which must be frequently checked.

Several criteria are used to estimate the external pressure that equalizes the internal pressure. The palpation, auscultatory, and oscillometric methods will be considered. They are known as "sphygmomanometric" methods (σφυγμός, pulse) because they measure the pulse pressure.

Palpation method of Riva-Rocci.[1] The arterial pulse is palpated below the place of compression. The pressure is gradually raised in the armlet until the pulse is no longer felt, and the pressure is read on the manometer. The pressure is raised a little more and then the air is allowed to escape gradually through the valve; when the pulse is again felt, the pressure is read. The average of the two readings is taken as the systolic pressure. The exact external pressure needed to suppress the pulse will be that which equalizes the maximum systolic pressure. With this method it is not possible to measure the diastolic pressure.

Auscultatory method. The sounds occurring simultaneously with the pulsation of the artery are explored below the site of compression. This method proposed by Korotkow is one of the most accurate of the nonsurgical methods, and it serves to measure the systolic and diastolic pressures. Air is pumped into the armlet until it is higher than the pressure in the artery. A stethoscope or a phonendoscope is placed over the brachial artery below the site of compression. As long as the pressure in the armlet is higher than the maximum systolic pressure the artery will be continuously collapsed and as no blood will pass below the compression, no sound will be heard. The pressure is gradually reduced by letting air escape from the armlet, until a soft sound is clearly heard at each heartbeat. The pressure is read on the manometer and then

[1] Riva-Rocci, S., *Gazz. med. di Torino*, **51**, 52, 1896.

the pressure is reduced slightly; the intensity of the sounds increases; the pressure is raised again gradually until the sounds are no longer heard. A second reading is made of the exact pressure at which the sounds disappear. The average of the two readings gives the systolic pressure. Air is gradually let out of the armlet, and as the pressure falls the sounds first reappear and then increase in intensity. The characteristics of the sounds vary from one subject to another, and it is not possible to make a systematic description, which will be valid for all cases, of the changes taking place as decompression progresses. At a definite moment in the process of decompression, the intensity of the sounds diminishes suddenly; this decrease is clear-cut and leaves the observer in no doubt as to when it occurs. The pressure read on the manometer at this moment is the diastolic pressure. If decompression is continued the intensity of the sounds diminishes gradually, and they disappear completely even before the decompression is completed. These changes in intensity have no definite significance. Arterial sounds are due to vibrations of the vascular walls and to the turbulent flow caused by the pressure wave modified by compression of the artery. A more precise explanation cannot yet be given.

Oscillometric method. This method is based on the analysis of the oscillations in the arterial wall (transmitted by the tissues and skin surrounding the artery) when different internal and external pressures act on it. The instruments must therefore have not only a pneumatic armlet and manometer, but also some device to make visible or register the oscillations. Pachon's[1] oscillometer (Fig. 95) is one of the instruments of this type most commonly used in France and other Latin countries.

The oscilloscope in Pachon's instrument (the wall of the aneroid *c*) is always submitted to the same internal and external pressure, whatever the pressure in the system. According to Marey's principle, therefore, it is always in optimum condition to respond with an ample oscillation to any difference between the external and internal pressures. When the connection *f* is open, changes in pressure caused by the pulse wave are transmitted equally to both sides of the capsule, and their effects are annulled. On closing *f* the pulsation is transmitted only to the aneroid, its wall being in optimum conditions for oscillations. The arterial wall, the tissues, and the pneumatic armlet are in a condition similar to the aneroid with respect to the capacity to oscillate. If the pressure in the armlet is constantly above that in the artery, the vessel will be permanently occluded and a small oscillation will be observed at each heartbeat (Fig. 96). These "supramaximal" oscillations are due to the water-hammer effect on the upper edge of the armlet produced by the column of blood in the occluded artery. If the pressure within and without the artery are the same, maximum oscillations will be observed; this is what happens when the external pressure is equivalent to the mean arterial pressure. Finally, when the external pressure is constantly below that in the artery, oscillations will be of

M E l P c b V S f a B

FIG. 95. Diagram of Pachon's oscillometer.

[1] PACHON, M. V., *Compt. rend. Soc. de biol.*, **66**, 733, 776, and 955, 1909.

very small amplitude or will disappear completely; this is what happens when the external pressure is lower than the diastolic pressure.

Pachon's instrument can be adapted to an oscillographic capsule, based on the same principle as the oscilloscope described, which will register the oscillations on a kymograph.

Pe

Decompression

Mn

Mx

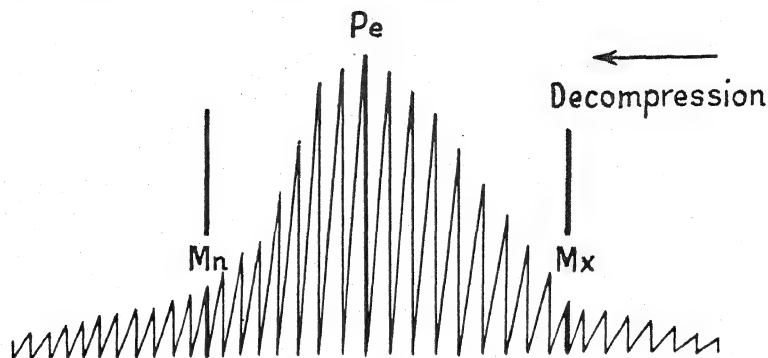



FIG. 96. Variations in amplitude of the oscillations of Pachon's oscillometer at different pressures. *Mx*, systolic pressure; *Mn*, diastolic pressure; *Pe*, oscillometric index.

There are several other instruments, oscilloscopes and oscillographs, which serve to measure human blood pressure; they are based on principles similar to those of Pachon's instrument. The main difficulty in all these instruments is that of establishing the exact moment at which a change in amplitude corresponds to one of the pressure values to be measured. Frequently the change in amplitude is not as definite as might be wished.

Oscillometric instruments are also very useful for exploring the arteries when occlusion by disease (*e.g.*, endarteritis obliterans) is suspected. A decrease in the amplitude or the absence of oscillations is a sign of partial or total arterial occlusion.

Accuracy of sphygmomanometric methods.[1] It is now a well-established fact that blood pressure cannot be measured with absolute accuracy by sphygmomanometers. Direct registration of blood pressure by means of intra-arterial manometers has shown that (*a*) even during normal breathing with slight sinus arrhythmia, systolic and diastolic pressure vary several millimeters of mercury from beat to beat, these variations increasing considerably when there is marked arrhythmia or deep breathing; (*b*) determinations of blood pressure in the brachial artery with the auscultatory method give figures 3 to 4 mm. Hg below the true pressure, with an average dispersion of ±8 mm. Hg; (*c*) diastolic pressure taken with the auscultatory method and read at the level when the intensity of the sound diminishes suddenly is more or less 8 mm. below the true value. In normal persons, therefore, an error of ±8 mm. Hg must be allowed for in single determinations. Sphygmomanometric methods, in spite of their shortcomings, are of great value in medical practice.

[1] BORDLEY, J., C. A. R. CONNOR, W. F. HAMILTON, W. G. KERR, and C. J. WIGGERS. *Circulation*. **4**, 503, 1951.

NORMAL VALUES OF BLOOD PRESSURE IN MAN

An immense number of determinations have been made to establish the normal values of blood pressure in man. Subjects of different sex, age, etc., candidates for life insurance, school children, university students, army recruits, etc., have been examined.

Individual differences are considerable. In groups of the same age, physical constitution, and occupation, one individual will be apparently normal with a systolic pressure of 100 mg. Hg or less, and another equally healthy will have a systolic pressure of 130 mm. Hg or even higher. All intermediate values between these extremes will be found. Average figures, therefore, have only a relative value. Even careful statistical studies in which the standard deviation or probable error is stated, although they are of interest in the general study of blood pressure, are not much help when deciding whether the blood pressure of a certain individual is normal or not.

The only legitimate conclusion is that the

blood pressure of an individual has a statistical probability of being normal. If, on the other hand, a man usually has a blood pressure of 130 mm. Hg when in a healthy condition, and is found to have a pressure of 95 mm. Hg, it must be considered as considerably below normal in his particular case, although statistically it has a certain probability of being normal.

The normal standard of blood pressure for an individual, as for other aspects of his constitution and functions, can be established by examining him in conditions of perfect health. If these data were noted on a health card it would be possible to detect any deviation from the normal as soon as it began.

The blood pressure has been correlated statistically with the age of the individuals. Systolic pressure in mm. Hg is roughly equal to 100 plus the age, except that after 50, figures above 150 mm. are certainly too high. Thus 120 mm. would be the normal blood pressure for a 20-year-old subject; 130 mm. for one 30 years old, etc. Such definite standards have not been established for diastolic pressure. It is usually admitted that diastolic pressure is half the systolic plus 10 to 20 mm. For example, a diastolic pressure of 70 to 80 should correspond to a systolic pressure of 120.

In Córdoba (Argentina) Moisset de Espanés found the following averages in 1,123 university students: systolic pressure, 130 mm. Hg, the extreme cases being 108 and 154; diastolic pressure, 78 mm. Hg, with 61 and 95 as extreme figures. In army recruits in training, many cases were observed with a systolic pressure between 90 and 100 mm. Hg.[1]

Wiggers[2] summarizes as follows the results obtained by several observers: "According to one group of observers the mean pressure values in mm. Hg are approximately 90/65 at three or four years, then increase rather steadily to about 100/70 at ten years, to 110/75 at fourteen years and to 115/75 at sixteen years of age. According to another group of observers the pressures are lower in early childhood, approximately 80/50 at the age of three years. Some believe that only one sudden jump occurs at the approximate age of ten or twelve years, others that another sudden increase is evident near the age of six years. Roughly, the pressures may be said to have increased approximately to 85/55 at six years, to 95/55 at ten years, to 115/65 at fourteen years and to 120/75 at sixteen years of age."

At birth direct measurements have given values of 80/46.

Determinations made with armlets 6 cm. wide give values that are too low. Armlets 12.5 cm. wide give more accurate results.

In old people systolic and diastolic pressures tend to decrease after 65 years of age. This is attributed to a progressively diminishing cardiac efficiency.

After puberty males have higher average blood pressures than females of the same age.

Orientals have lower blood pressure averages than European or American whites.

During sleep, except when dreams disturb normal conditions, systolic pressure falls 15 to 30 mm., and diastolic pressure 5 to 10 mm.

Posture produces changes which vary in different individuals. Physical exercise produces a rapid increase at the beginning and then a slight fall, with stabilization at a level above the resting figure. The type, intensity, and duration of the exercise, and the condition (training, etc.) of the subject have considerable influence on the response of the blood pressure (see Chap. 47).

[1] Moisset de Espanés, E., *Rev. Soc. argent. de biol.*, **17**, 564, 1941.

[2] Wiggers, C. J., "Physiology in Health and Disease," 3d ed., Lea & Febiger, Philadelphia, 1939, p. 646.

BIBLIOGRAPHY

Alberti, V. A., R. González-Segura, and A. Lanari, *Medicina, Buenos Aires*, **4**, 412, 1944 (respiratory variations in blood pressure).

Alvarez, W., *Arch. Int. Med.*, **32**, 17, 1933 (arterial blood pressure in man).

Battro, A., R. González-Segura, C. A. Eliçabe, and E. Araya, *Ann. Inst. Fis. Aplic. a la Pa!ol. hum.*, **4**, 39, 1942 (arterial blood pressure and respiration).

Battro, A., R. González-Segura, C. A. Eliçabe, and E. Araya, *Arch. Int. Med.*, **73**, 29, 1944 (records with Hamilton's manometer in man).

Bellis, C. J., *Surgery*, **13**, 27, 1943 (measurements of velocity of blood flow).

Castellanos, A., R. Pereiras, and P. P. de los Reyes, Angiocardiography and Its Value, *Am. Heart J.*, **25**, 298, 1944.

Hess, W. R., in Bethe's "Handbuch der normalen und pathologischen Physiologie," **7**, 919, Springer, Berlin, 1927 (hydrostatic laws and hydrodynamics).

Mattes, K., "Kreislaufuntersuchungen am Menschen mit fortlaufend registrierenden Methoden," Thieme, Stuttgart, 1951.

Norris, G. W., H. C. Bazett, and T. M. McMillan,

"Blood Pressure, Its Clinical Applications," Lea & Febiger, Philadelphia, 1927.

PLESH, J., "Blood Pressure and Its Disorders," 2d ed., Williams & Wilkins, Baltimore, 1947.

RAGAN, C., and J. BORDLEY, *Bull. Johns Hopkins Hosp.*, **69**, 504, 1941 (accuracy of methods for measuring blood pressure).

VON RECKLINGHAUSEN, H., "Blutdruckmessung und Kreislauf in den Arterien des Menschen," T Steinkopf, Leipzig, 1940.

WESLER, K., and A. BÖGER, *Ergebn. d. Physiol.*, **41**, 292, 1939 (dynamics of the arterial system).

WIGGERS, C. J., "Modern Aspects of Circulation in Health and Disease," Lea & Febiger, Philadelphia, 1923.

WIGGERS, C. J., "The Pressure Pulses in the Cardiovascular System," Longmans, New York, 1928.

ZWEIFACH, B. W., and E. SHORR, "Factors Regulating Blood Pressure," Josiah Macy, Jr., Foundation, New York, 1949.

CHAPTER 20

Vasomotor Activity

SMALL ARTERIES AND arterioles are provided with a thick layer of smooth muscle, which is under nervous control. Contraction and relaxation of this muscular layer determine the diameter of the arteries. Substances in the blood may also modify the diameter of the blood vessels, either by direct action on the arterial muscles or by modifying the activity of the vasomotor centers. All these agents act in a coordinated way, as part of integrative mechanisms that assure the unity and constancy of the organism. They contribute to maintain an adequate arterial blood pressure and to distribute the blood to different regions according to the physiologic needs of the moment.

VASOMOTOR NERVES

Claude Bernard[1] discovered the vasomotor nerves in 1851. He observed that cutting the cervical sympathetic in a rabbit caused dilatation of the arteries in the territory it innervated, clearly visible in the ear on the denervated side.

The vasomotor centers continuously discharge impulses along the sympathetic nerve fibers which keep the arterial muscles in a state of tonic contraction. When the sympathetic is cut, these impulses are interrupted and the arteries and arterioles dilate passively.

Electrical stimulation of the peripheral end of the cut sympathetic produces vasoconstriction (Brown-Séquard, 1852). The blood vessels, which are dilated after the nerve has been cut, contract on stimulation, and only the larger ones appear as thin filaments to the naked eye. There is marked contrast between the vascularization of the two ears, but in the opposite direction to that observed in Claude Bernard's experiment. Arterial constriction persists as long as the nerve is stimulated, and afterward disappears gradually. This experiment also shows that the tonic constriction of the arteries maintained by nerve impulses can be increased by stronger stimulation of the nerve fibers.

[1] BERNARD, C., *Compt. rend. Soc. de biol.*, **40**, 168, 1852.

Claude Bernard[1] also discovered the vasodilator nerves. He observed an increase in the blood flow through the submaxillary salivary glands when the chorda tympani was stimulated.

Stimulation of a nerve can provoke vasoconstriction, vasodilatation, or both. The cervical sympathetic and splanchnic nerves are vasoconstrictor nerves; the splanchnics are the most important of all vasoconstrictor nerves, because of the large territory they innervate and the intensity of their effect. The chorda tympani and the nervi erigentes (pelvic nerves) are vasodilators. Stimulation of the majority of the large nerve trunks may produce either vasoconstriction or vasodilatation, as they carry vasoconstrictor and vasodilator fibers. The origin of these fibers, although complicated, can be determined, as will be demonstrated further on.

Effects of arterial dilatation and constriction. Vasodilatation in an organ causes it to redden, its volume and temperature increase, and the blood flow through it is considerably greater. Any lesion that cuts a blood vessel causes abundant hemorrhage. The increase in blood flow is relatively greater than the increase in the vascular bed; therefore the blood flows more rapidly, especially through the capillaries. The quick passage of the blood gives less time for the exchange of oxygen and CO_2 between the blood and tissues; therefore the arteriovenous difference for both gases diminishes.

Since arterial dilatation causes a decrease in peripheral resistance, the arterial blood pressure

[1] *Ibid.*, **46**, 159, 1858.

falls and the venous pressure rises. The fall in arterial pressure is less marked than the rise in venous pressure, except when arterial dilatation extends to a large vascular territory. In this case venous pressure can also fall because of the stasis caused by the increase in the vascular bed.

wax, or plasticine. The organ must not fill the plethysmograph completely; there must be an air space to allow for its expansion. The air space is connected by means of rubber tubing with a Marey tambour (for registering on a kymograph) or a Frank segment capsule (for optical registration). The changes in volume in the organ will modify the pressure in the air space, and a record called a plethysmogram will be obtained.

V
A
U

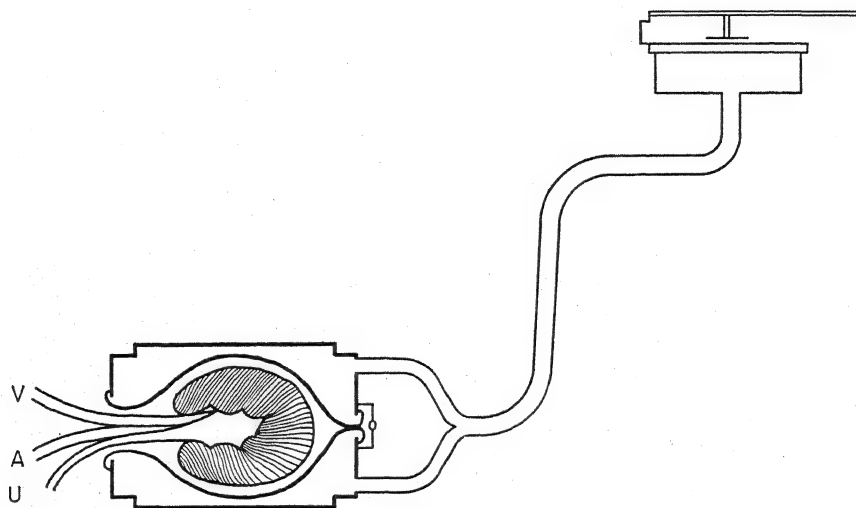



FIG. 97. Plethysmograph for registering the volume of the kidney. *A*, renal artery; *V*, renal vein; *U*, ureter.

Vasoconstriction produces the opposite effects.

The best criteria for measuring the degree of vasodilatation or vasoconstriction are (*a*) the changes in volume of the organ and (*b*) the blood flow through the efferent veins. The former is most frequently used, and a plethysmographic record is obtained. It must not be forgotten that changes in volume and blood flow can be produced passively by an increase in the blood pressure, without the active participation of the arterial muscles; blood pressure must, therefore, be recorded simultaneously in all cases.

Plethysmography. This consists in recording the changes in the volume of an organ. The organ is enclosed in a box with rigid walls, the form of which varies according to the organ examined. The instrument is known as a plethysmograph (Fig. 97).

A plethysmograph opens on hinges as if it were the two valves of a clam. The free edges fit closely except at one place, where an orifice leaves a passage for the nerves, vessels, and excretory ducts of the organ; it must be sufficiently ample not to constrict these structures, but not so wide as to leave the inside of the plethysmograph in communication with the outside. The opening is closed with petroleum jelly, wax, or plasticine.

The plethysmograph so far described is used for pediculated organs (such as the kidney); they are also known as "oncometers" (Greek ὄγκος, bulk, and μέτρον, measure). When the volume of a whole limb is to be measured, the limb is introduced into a cylinder (somewhat larger than the limb) open only at one end (Fig. 98). The cardiometer (Fig. 24) described in Chap. 12 is a plethysmograph.

Photoelectric plethysmography consists in transillumination of an organ (preferably with ultraviolet light) and registering, by means of a photoelectric cell, variations in the light passing through the organ.[1]

With a sufficiently sensitive plethysmograph, a rise and fall in volume at each heartbeat will be observed. This pulsation is due to the expansion of the blood vessels produced by the increase in pressure and flow caused by the systolic discharge. In addition to these "cardiac" fluctuations in volume, other much slower changes

[1] Matthes, K., and F. Gross, *Arch. exper. Pathol. u. Pharmakol.*, **191**, 391, 1938.

are observed, which are due to (*a*) changes in arterial blood pressure, which passively modify the expansion of the blood vessels in the organ; (*b*) active modifications in the diameter of the blood vessels, caused by contraction or relaxation of the arterial muscles. A simultaneous blood-pressure record will permit these changes to be attributed to one cause or the other. Figure 99 reproduces a record of arterial blood pressure, obtained with a mercury manometer, and a plethysmogram of a leg of a dog that had received an intravenous injection of adrenaline. The blood pressure rose, and the volume of the leg increased for a short period and afterward decreased. This decrease is due to vasoconstriction; otherwise the rise in blood pressure would produce a passive increase in volume. The increase in blood pressure is due to the increase in peripheral resistance brought about by the constriction of the arteries.

Origin and distribution of vasoconstrictor fibers. Langley's[1] classic experiments on the effects of nicotinization of nerve ganglia demonstrated the distribution and principal connections of vasoconstrictor fibers. Almost all these fibers belong to the thoracolumbar division of the autonomic nervous system. They emerge from the spinal cord in both ventral roots of the twelve thoracic nerves and the two or three upper lumbar nerves. These fibers form part of the white rami that go from the spinal nerves to the ganglia of the homolateral sympathetic chain, where they end synaptically on the ganglion cells. The axons of these neurons (postganglionic fibers) form part of the gray rami going back to the spinal nerves, with which they are distributed to the tissues.

FIG. 98. Plethysmograph for registering the volume of a limb.

Nicotine suppresses the effects of preganglionic stimulation, but postganglionic stimulation continues to produce results. This experiment shows that the fibers are interrupted in the ganglia.

Vasoconstrictor fibers for the head leave the spinal cord in the first four thoracic nerves and join the sympathetic chain. Some have their relay station in the thoracic ganglia and others

[1] LANGLEY, J. N., *J. Physiol.*, **11**, 123, 1890.

in the cervical ganglia. Finally they reach their destination through the periarterial plexuses. These were the fibers discovered by Claude Bernard in the cervical sympathetic.

Vasoconstrictor fibers for the upper limb leave the spinal cord in the fourth to the tenth thoracic nerves.

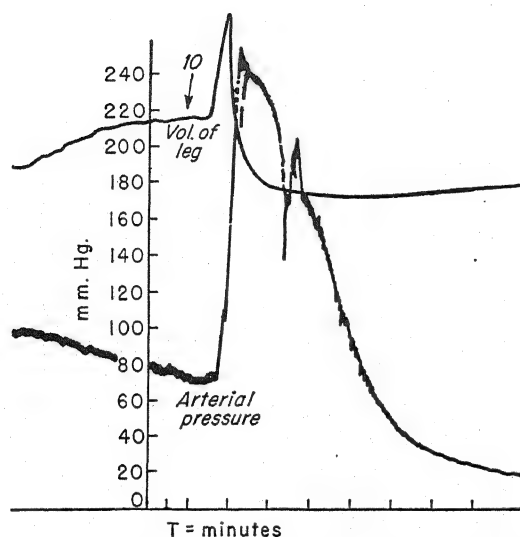



FIG. 99. Plethysmogram of the leg and arterial blood pressure of a dog injected intravenously (arrow) with adrenaline. The volume of the leg after an initial dilatation diminishes considerably because of vasoconstriction; vasoconstriction causes the rise in blood pressure.

The splanchnic area, which extends to the abdomen and pelvis, receives vasoconstrictor fibers from the last five thoracic and the first two or three lumbar nerves. The majority of these fibers form part of the major and minor splanchnic nerves.

Vasoconstrictor fibers for the lower limb leave the spinal cord in the second and third lumbar nerves.

The thoracic and abdominal walls receive vasoconstrictor fibers from the corresponding spinal segment. There is a strictly homolateral innervation, although there are numerous anastomoses between the blood vessels across the mid-line of the body.[1]

Constrictor fibers for the coronary vessels form part of the vagi nerves, which belong to the cranial division of the parasympathetic.

Origin and distribution of vasodilator fibers. The distribution of vasodilator fibers does not follow such a simple plan as that of the vasoconstrictors. Vasodilator fibers may be classified in three main groups: (*a*) parasympathetic; (*b*) sympathetic; (*c*) the so-called "antidromic" vasodilators.

Parasympathetic vasodilators belong to the cranial and sacral divisions of the parasympathetic. They exert no tonic action, but on stimulation they provoke vasodilatation in the territory they innervate. Among the cranial fibers are those discovered by Claude Bernard in the chorda tympani for the submaxillary salivary gland; these, together with those which emerge from the central nervous system in the seventh nerve, join the major superficial petrosal nerve and end in the sphenopalatine ganglion, whence the postganglionic fibers go to the lachrymal, nasal, and buccal glands. Vasodilator fibers for the tongue leave the seventh nerve and join the lingual nerve. The ninth nerve (glossopharyngeal) sends fibers through the minor superficial pretosal nerve to the parotid gland. The sacral division gives the nervi erigentes to the genital organs.

Sympathetic vasodilators produce, in experimental conditions, less clearly marked effects, because they are masked by the simultaneous stimulation of the predominant vasoconstrictors. For vasodilatation to be evident, the vasoconstrictors must first be paralyzed by an adequate dose of ergotamine.

The effect of sympathetic vasodilators is most clearly seen in the coronary arteries, because here the vasoconstrictors form part of the parasympathetic and not of the sympathetic division of the autonomic nervous system. There are sympathetic vasodilators for the skeletal muscles, the importance of which varies in the different species. There are sympathetic vasodilators in the splanchnic nerves: after the injection of large doses of ergotoxin, stimulation of the splanchnic nerve causes a fall in blood pressure[1] (for details see Burn[2]).

Antidromic vasodilators are exceptional in the organization of the nervous system because they form part of the dorsal roots of the spinal nerves: peripheral stimulation of these nerves provokes a motor effect, *i.e.*, relaxation of the muscles in the arterial walls. This is contrary to the Bell-Magendie law, which states that the ventral roots are motor and the dorsal roots sensory;

[1] GILDING, H. P., *J. Physiol.*, **74**, 34, 1932.

[1] DALE, H. H., *J. Physiol.*, **46**, 294, 1913.
[2] BURN, J. H., *Physiol. Rev.*, **18**, 137, 1938.

they are called antidromic vasodilators because they conduct impulses in the opposite direction to the majority of the fibers traveling in the dorsal root (see Chap. 69, Antidromic Conduction).

Stricker[1] observed that peripheral stimulation of the dorsal roots provoked vasodilatation in

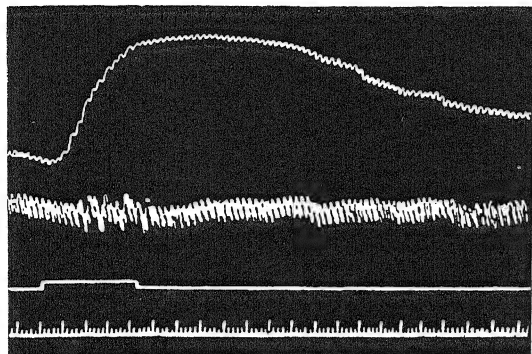



FIG. 100. Stimulation of the peripheral end of the seventh lumbar dorsal root in the dog. From above down: volume of hind leg, arterial blood pressure, signal marking stimulation, and time in seconds. (*Bayliss, W. M., J. Physiol., vol. 26, p 173, 1901.*)

the corresponding cutaneous and visceral territories. This fact was later confirmed by Bayliss[2] (Fig. 100). The origin and nature of the fibers concerned are still uncetains. A widely accepted interpretation is that they are sensory fibers, belonging to the unipolar sensory neurons of the spinal ganglia. Some of the branches arise as sensory (pain) fibers in the skin and tissues; other branches of the same fibers arise in the arterial walls. Stimulation of these fibers, either between the ganglion and the spinal cord or peripherally to the ganglion, spreads throughout the whole neuron and all its branches. It therefore spreads also to those arising in the arteries, where it provokes relaxation of the muscle layer. Antidromic vasodilatation, however, is not a mere experimental curiosity without functional significance; certain cutaneous vascular responses have been explained by the existence of the fibers and their peculiar distribution. For example, mechanical irritation of the skin or the application of mustard oil provokes local cutaneous vasodilatation, but this does not occur if the skin has been anesthetized with cocaine. Nerve conduction is therefore necessary for this type of reaction, which appears to be a reflex; but it is not a true reflex because section of the root, centrally or peripherally to the spinal ganglion, does not suppress the response. A so-called "axon reflex" seems to have been provoked, *i.e.*, mechanical irritation or mustard oil stimulates the cutaneous branches of the axon and the impulses spread to all its branches traveling centrifugally to the artery (antidromic conduction) where they provoke relaxation (Fig. 101).

Herpes zoster, produced by inflammation of the spinal ganglion, has been attributed to this type of mechanism.

Antidromic vasodilatation has also been attributed to the stimulation of motor fibers, which would emerge from the spinal cord in the dorsal instead of in the ventral roots.

Dissociation of the vasomotor effects produced by stimulation of mixed nerves. Usually nerve trunks have vasodilator and vasoconstrictor fibers. According to the nature of the stimulus or the condition of the nerve, the

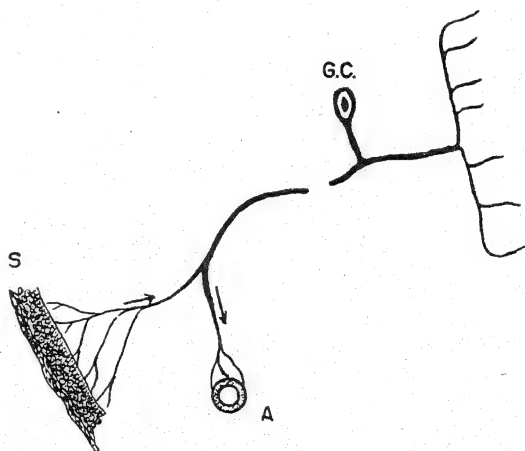



FIG. 101. Diagram of vasodilatation produced by antidromic impulses. *GC*, dorsal root ganglion cell; *S*, skin; *A*, arteriole. The nervous fiber arises in the skin and in the arteriole. The arrows mark the direction in which the nerve impulses spread when the skin is stimulated, and show how a pseudoreflex can be produced even after the root has been cut.

result of stimulation will be constriction or dilatation. Constriction is usually the predominant effect, and vasodilatation is not easily observed.

Stimulation of the peripheral end of a mixed nerve, immediately after it has been cut, pro-

[1] STRICKER, S., *Wien. Sitzber.*, **74**, 173, 1876.
[2] BAYLISS, W. M., *J. Physiol.* **26**, 173, 1901; **28**, 276, 1902.

duces marked vasoconstriction. Stimulation of the same nerve 4 days after it has been cut has the opposite effect. Constrictor fibers therefore degenerate more rapidly than dilator ones. Vasodilator fibers are also more resistant to cold than the constrictors. Stimulation at rates between 16 and 64 induction shocks per second provokes vasoconstriction; at rates of 1 to 4 shocks per second there is vasodilatation. After injection of ergotamine, which paralyzes the vasoconstrictor fibers, stimulation produces vasodilatation.

Mechanism of vasodilatation and constriction. Vasoconstrictor fibers exert a continuous "tonic" effect. Vasodilatation can be provoked by suppressing this effect (passive vasodilatation) and, in a more intense degree, by stimulation of vasodilator fibers (active vasodilatation). Vasodilator fibers exert no "tonic," permanent action. There is a certain degree of vasoconstriction permanently maintained in all the arteries; vasodilatation is transitory and localized to a definite territory.

Chemical mediators. As in other aspects of visceral innervation, the vasomotor nerves produce their effects by the release of chemical mediators. Vasomotor fibers are cholinergic or adrenergic according to whether they liberate acetylcholine or sympathin.

Not all the vasomotor fibers can as yet be classified as belonging to one or another of these types. Sympathin is liberated by vasoconstrictor fibers, with the exception of those acting on the coronary arteries. These fibers are adrenergic vasoconstrictors and form part of the sympathetic system. Cholinergic vasoconstrictors are exceptional. The vasoconstrictors of the coronary arteries belong to this type; they form part of the parasympathetic system.

Cholinergic vasodilators include

1. All parasympathetic vasodilators (chorda tympani, petrosal nerves, nervi erigentes, etc.).
2. Antidromic vasodilators. These fibers also release histamine, according to Lewis and Marvin.[1]
3. Sympathetic vasodilators of skeletal muscles. These, together with the secretory fibers of the sweat glands, are a notable exception to the general rule that sympathetic postganglionic fibers are adrenergic.

[1] Lewis, T., and H. M. Marvin. *Heart*, **14**, 27, 1927.

Adrenergic vasodilators dilate the coronary arteries; they belong to the sympathetic system.

Summary. Vasoconstrictors, with the exception of those of the coronary arteries, are adrenergic; vasodilators are cholinergic, with the exception of those of the coronary arteries. The effect on the arteries seems to depend more on the nature of the chemical mediator than on the origin of the nerve fibers. For example, a discharge of adrenaline from the adrenals will provoke generalized vasoconstriction, except in the coronary arteries. These arteries will therefore dilate in this condition that demands a greater effort from the heart to overcome the increased peripheral resistance and provide the myocardium with an increased blood flow.

Ergotamine paralyzes adrenergic fibers, and atropine paralyzes cholinergic fibers. Atropine has less effect on the vasodilators than on the cardiac vagal nerves and the secretory fibers of the chorda tympani. Much larger doses are needed to suppress vasodilatation than those needed to suppress cardiac inhibition or salivary secretion.

VASOMOTOR CENTERS

The activity of vasomotor fibers is coordinated by nerve centers. The principal vasomotor center, situated in the medulla, was discovered by Ludwig. There are subsidiary centers in other parts of the central nervous system. The center in the medulla has a permanent vasoconstrictor effect, which is of great importance to the maintenance of the arterial blood pressure. Complete section of the spinal cord just below the medulla produces a fall in blood pressure, which causes a condition of circulatory failure known as shock. This hypotension, with stasis due to widening of the vascular bed, is not permanent. Approximately four hours after the section (during which time the animal is kept alive by artificial respiration because the paths from the respiratory center have been severed) the blood pressure begins to rise and can reach a normal, or approximately normal, level. Recovery is due to an increased activity of the subsidiary spinal centers, which compensate, up to a certain point, for the loss of the main vasomotor center. If now the spinal cord is destroyed there is a second, and more serious, fall in blood pressure. After a time the blood pressure again rises, because a local automatic tonus develops in the arterial muscles.

Sections of the nerve stem above the medulla have little effect on the blood pressure and do not produce any fundamental disturbance in vasomotor reflexes. This well-known fact has been confirmed and amplified by Braun Menéndez. Stimulation of the hypothalamus provokes vascular reactions, and the cerebral cortex sends impulses that act on the arterial muscles in certain types of response, such as blushing, but these higher centers do not normally play a part in the maintenance of vascular tonus or in the fundamental vascular reflexes.

Cocainization, freezing, or destruction of the floor of the fourth ventricle on both sides of the mid-line produces the same effects as a transverse section of the spinal cord just below the medulla. Stimulation of the tovea on both sides of the mid-line provokes hypertension, and stimulation of the obex near the calamus scriptorius produces hypotension. It is doubtful whether there is a true vasodilator center.

The Mechanism of Vasomotor Activity

The permanent activity of the vasomotor center is due to the constant play of diverse influences. It is submitted to the action of agents brought to the center by the blood, to the products of its own metabolism, and to nerve impulses of multiple origin.

Humoral agents. CO_2 is the substance brought by the blood that has the most definite and well-demonstrated effect on the vasomotor center. Increased CO_2 in the blood causes general vasoconstriction; all the tissues become pale and the blood pressure rises. This response is more marked if both vagus nerves have been cut, because of the suppression of cardio-inhibitory reflexes that control hypertension; in this case the blood pressure may reach a level as high as twice the normal value (Fig. 102). Hypertension of asphyxia is due to accumulation of CO_2 in the blood. In extreme cases other metabolic products (*e.g.*, lactic acid) and anoxia may also stimulate the vasomotor center. The constrictor effect of CO_2 as described is due to its action on the center and not on peripheral structures; this can be clearly demonstrated by perfusing the head of an animal with hypercapnic (excess of CO_2) blood. The head of a dog *B* (Fig. 103) receives blood from another dog *A*. Cross circulation is maintained by anastomosing the carotid arteries and jugular veins of both animals.

A decrease in the CO_2 content of the blood causes a decrease in the activity of the vasomotor center, with general vasodilatation and hypotension. In extreme cases a state of shock may develop.

Changes in the chemical composition of the blood can also act indirectly on the vasomotor

Fig. 102. Effect of asphyxia on arterial blood pressure (registered with mercury manometer) of a curarized cat with both vagi cut. 1, artificial respiration is stopped; 2, artificial respiration is renewed; 3, the animal breathes in an atmosphere of nitrogen; 4, air is restored. (*Mathison, G. C., J. Physiol., vol. 41, p. 416, 1910.*)

center by stimulating chemoreceptors, such as the aortic and carotid bodies; in this case impulses are conducted to the center along afferent nerves.

Nervous influences. The activity of the vasomotor center is also under the influence of nervous impulses, which arrive from many parts of the organism and thus modify reflexly the diameter of the blood vessels. Somatic, visceral, and vascular reflexes will be considered.

Somatic reflexes. Stimulation of the central end of any somatic sensory nerve causes reflex changes in the diameter of the blood vessels. For example, electrical stimulation of the central end of the sciatic nerve provokes arterial hypotension or hypertension, according to the strength, frequency, and duration of the stimulus. Usually moderate stimulation causes hypertension (vasoconstriction) and strong stimulation causes hypotension (vasodilatation). Pain of moderate intensity causes vasoconstriction and paleness; intense pain causes vasodilatation, which is sometimes so marked as to bring about a state of shock.

Cold produces vasoconstriction in the skin, and heat produces vasodilatation. These re-

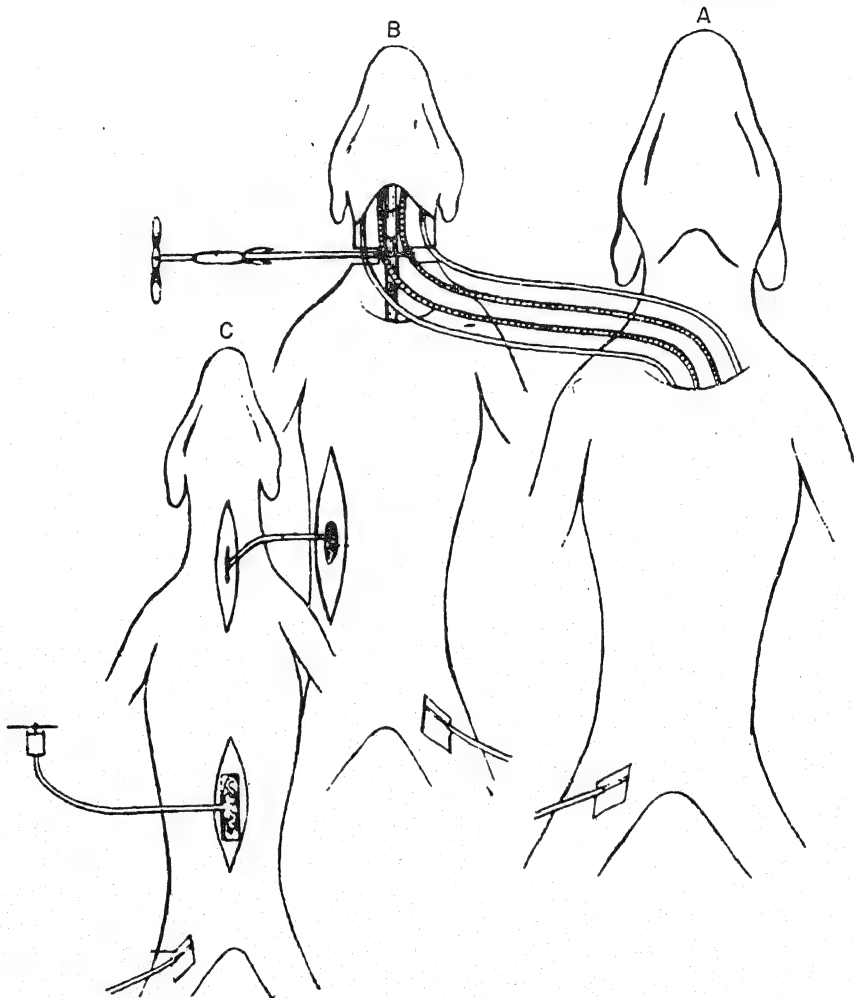



FIG. 103. Diagram of crossed-circulation experiment. The head of dog *B* is joined to the trunk only by the spinal cord; it receives its blood supply from dog *A*. Dog *C* is adrenalectomized; it receives blood from the adrenal of dog *B*. (*After Heymans, C., Arch. internat. de pharmacodyn. et de thérap., vol. 35, p. 269, 1929.*)

sponses form part of the mechanism that regulates the body temperature; they are accompanied by simultaneous vascular reactions in the splanchnic area, but in the opposite direction.

Visceral reflexes. Stimulation of the afferent fibers of visceral nerves (compression of the testes, pulling on the mesentery, etc.) provokes vascular reactions. The type of response is conditioned by the strength and other qualities of the stimulus.

Vascular reflexes. These are of great importance in the maintenance of arterial blood pressure. They are produced by impulses that arrive at the vasomotor center from receptors (pressoreceptors) situated in certain parts of the vascular system and travel along special paths, described when considering the nervous control of cardiac activity (see Chap. 17). The cardioaortic (de Cyon's depressor nerve) and Hering's nerves, which innervate the aorta and carotid sinus and are stimulated by a rise in arterial pressure, carry impulses not only to the cardioinhibitory center but also to the vasomotor center, which is inhibited, thus causing generalized vasodilatation, the intensity of which is proportional to the rise in pressure. Excitation of the cardioinhibitory center and inhibition of the vasomotor center combine in provoking a fall in blood pressure and thereby a return to the normal level. When the blood pressure falls below normal, the opposite reflex effects are produced, *i.e.*, tachycardia and vasoconstriction, and the blood pressure therefore rises.

Impulses from the vascular pressoreceptors are also carried to the respiratory center and provoke respiratory reflexes (see Chap. 32).

The nerve fibers innervating the aortic and carotid bodies carry impulses originated in chemoreceptors, sensitive to changes in the chemical composition of the blood. An increase in CO_2, a decrease in the pH of the blood, and anoxia (diminished oxygen tension) stimulate the chemoreceptors and provoke hypertension due to vasoconstriction (stimulation of the vasomotor center) and tachycardia (inhibition of the cardioinhibitory center) (Fig. 104). Several drugs injected into the blood also stimulate the chemoreceptors.

The aortic arch and the carotid sinus are the most important vasosensory areas for the regulation of the blood pressure. The pacinian corpuscles on the mesenteric arteries also discharge afferent impulses in relation to the degree of distention of the arterial walls, but the reflexes that these impulses provoke are not yet well known.

Drugs with inhibitory effects on the vasomotor nervous system.[1] Several drugs inhibit the transmission of vasomotor impulses along the sympathetic. They are useful auxiliaries in the treatment of certain diseases of the circulation, *e.g.*, arterial hypertension, and in the study of the pathogenesis of these diseases; they are also valuable in certain surgical procedures. These drugs may be classified into three main groups: (*a*) agents with central activity which block vasomotor reflexes (cardioaortic, carotid sinus) at the vasomotor centers, *e.g.*, *plasmochin*,[2] *pentachin*,[3] and the central blocking component of ergot;[4] (*b*) antiadrenergic agents which block sympathetic impulses at the periphery and inhibit the effects of adrenaline and noradrenaline, *e.g.*, *dibenamine*,[5] and priscol (benzylimidazoline);[6] (*c*) agents which block sympathetic and parasympathetic ganglia and which not only do not inhibit pressor effects of adrenaline and noradrenaline in animals and man but reinforce them; the best known is tetramethylammonium chloride.[7] Derivatives of polymethylene-bis-trimethylammonium[8] also have interesting pharmacological properties. The C^{10} compound, known as decamethonium, has curarizing activity; the C^5 and C^6 compounds, called *pentamethonium* and *hexamethonium* respectively, exert a blocking action on the ganglia, which in man lasts longer and has fewer side effects than other drugs used for this purpose. Good results have been reported in the treatment of certain cases of arterial hypertension.[9]

[1] Finnerty, F. A., Jr., and E. D. Freis, *Circulation*, **2**, 828, 1950.

[2] Moe, G. K., and M. H. Leevers, *Federation Proc.*, **5**, 193, 1946.

[3] Richardson, A. P., H. A. Walker, and B. S. Miller, *Proc. Soc. Exper. Biol. & Med.*, **65**, 258, 1947.

[4] Rothlin, E., *Bull. schweiz. Akad. d. med. Wissensch.*, **2**, 249, 1947.

[5] Nickerson, M., and G. M. Nogamuchi, *J. Pharmacol. & Exper. Therap.*, **93**, 40, 1948.

[6] Chezz, D., and F. F. Yonkman, *Proc. Soc. Exper. Biol. & Med.*, **61**, 127, 1946.

[7] Acheson, G. H., and G. K. Moe, *J. Pharmacol. & Exper. Therap.*, **87**, 220, 1946; Freis, E. D., C. J. MacKay, and W. Oliver, *Circulation*, **3**, 254, 1951.

[8] Paton, W. D. M., and E. J. Zaimis, *Nature*, **162**, 810 1948.

[9] Smirk, F. H., and K. S. Alstad, *Brit. M. J.*, **1**, 121, 1951.

Humoral Agents That Act Directly on the Arteries

Vasomotor nerves release chemical mediators that act on the arterial muscles. Nerve action is ultimately the result of a chemical action; it is not surprising, therefore, that chemical substances in the blood may produce contraction and relaxation of the arterial muscles. These substances are (*a*) products of tissue metabolism; (*b*) hormones produced by the endocrine glands.

Products of tissue metabolism that act directly on the arterial muscles usually provoke vasodilatation. CO_2, lactic acid, other acids, adenylpyrophosphate, etc., are some of the principal substances that act in this way. Probably they are of importance in the production of the local vasodilatation that occurs when a tissue increases its activity. Increased activity causes greater production of these metabolic substances. There is considerable vasodilatation in active muscles, which is probably caused by this mechanism.

Hormones play a considerable and complex part in the regulation of arterial diameter and therefore of the blood pressure. Some hormones, *e.g.*, adrenaline, secreted by the adrenal medulla, and vasopressin, produced by the neurohypophysis, have a direct and marked effect. Adrenaline has a vasoconstrictor effect on all the arteries with the exception of those of the heart. It is active in cases of emergency when there is a disturbance in the functional equilibrium, but it is not indispensable for the maintenance of arterial tone and normal blood pressure.

Vasopressin also has a general vasoconstrictor

Fig. 104. Principal mechanisms that control the tonus of the blood vessels. *VMC*, vasomotor center; *AxR*, axon reflex; *PSD*, parasympathetic vasodilators; *DRD*, dorsal root dilators; *SC*, sympathetic vasoconstrictors. (*After Heymans, C. and J. J. Bouckaert, Ergebn. d. Physiol., vol. 41, p. 28, 1939.*)

activity, but it is less marked in the cerebral, renal, and pulmonary areas. It is not yet possible to say how important vasopressin is for the maintenance of arterial blood pressure, except in amphibians, in which the part it plays in maintaining vascular tone and blood pressure has been demonstrated.[1]

Kalikrein and vagotonin are vasodilating substances obtained from pancreatic extracts, but their hormonal nature has not been demonstrated.

The influence of the endocrine glands on the diameter of the blood vessels and blood pressure will be considered in the chapters on each gland.

The principal factors that take part in the maintenance and regulation of blood pressure are represented diagrammatically in Fig. 104.

Hypertensin (angiotonin). Renal ischemia produced by partial occlusion of the renal artery[2] is due to the release into the blood stream of a specific substance.[3] The work of Braun Menéndez *et al.*[4] and of Page *et al.*[4] has shown that the ischemic kidney discharges renin, a protein that acts as an enzyme and transforms an α globulin in the blood (known as "hypertensinogen") into a powerful vasoconstrictor, hypertensin (called "angiotonin" by Page) (Fig. 94). This mechanism probably plays a part in some cases of human hypertension. Renin is discharged by the kidney and hypertensin is formed when there is a marked fall in blood pressure (hemorrhage, anaphylactic shock); therefore this mechanism probably plays a physiologic part in the recovery of normal blood pressure in these conditions (see "Arterial hypertension of renal origin," Chap. 63).

Role of arterial contractility. The muscles of the artery are an important factor in the maintenance and regulation of blood pressure; they also regulate the circulation through the different vascular territories and organs. In this last respect they are a factor in the circulatory balance between the somatic and splanchnic areas in muscular exercise, in the regulation of body temperature, in renal and other secretions, and in the water balance between the blood and tissues.

Effect of sympathectomy on vasomotor tone. Surgical removal of segments of the sympathetic

[1] Orías, O., *Rev. Soc. argent. de biol.*, **10**, 91, 1934.

[2] Goldblatt, H., *Ann. Int. Med.*, **11**, 69, 1937; *Harvey Lect.*, **33**, 237, 1938.

[3] Houssay, B. A., *et al.*, *J. Physiol.*, **94**, 281, 1938.

[4] See bibliography at the end of the chapter, papers by Braun Menéndez *et al.* and Page *et al.*

ganglion chain produces vasodilatation in the territory innervated by the fibers which are thus disconnected from the centers. Sympathectomy has been performed in human subjects in order to relieve arterial hypertension. A certain degree of vasomotor tone is, however, recovered after a time, and it has been postulated that ganglionectomy does not sever all vasoconstrictor fibers.[1] Moreover, in some patients autochthonous vascular resistance that is not of nervous origin may occur after sympathectomy, but so far there is no satisfactory explanation for this development.[2]

Arterial hyperten ion. A blood pressure above the normal is frequently observed in adults of both sexes. In some cases it is possible to attribute this hypertension to a definite pathologic condition of which it is a sign, *e.g.*, in many acute and chronic renal diseases. In other cases, even careful and detailed examination reveals no other disturbance except hypertension. This so-called "essential hypertension" is one of the major problems of medicine that still awaits a solution. Its cause is unknown, and the mechanisms that take part in it are insufficiently understood; treatment of the disease is therefore symptomatic and empirical. A recent important advance has been the discovery of the mechanism of hypertension due to renal ischemia (renin-hypertensin). Experimental, more or less permanent, hypertension has been obtained by cutting the cardioaortic and Hering's nerves on both sides. Hypertension has also been obtained experimentally by injecting kaolin into the cisterna magna, thus provoking an increase in cranial pressure. The latter two types of hypertension differ in several ways from human cases of hypertension.

Wiggers[3] has made a careful examination of the mechanism of production of hypertension, examining diastolic, systolic, and pulse pressures and the shape of the pulse waves. Three factors can enter into play: (*a*) increase in heart rate and systolic discharge; (*b*) decrease in caliber and elasticity of the great arteries; (*c*) increase in resistance at the level of the arterioles. The first factor can be eliminated, because many observations have shown that the minute volume and systolic discharge are within normal ranges in patients with hypertension. Peripheral resistance, on the other hand, is increased, and there are signs of diminished

[1] Randal, W. C., W. F. Alexander, A. B. Hertzman, J. W. Cox, and W. P. Henderson, *Am. J. Physiol.*, **160**, 441, 1950.

[2] Mebdlowitz, M., and A. S. W. Touroff, *Circulation*, **5**, 577, 1952.

[3] Wiggers, C. J., *Am. Heart J.*, **16**, 515, 1938.

distensibility of the aorta. There is no direct relation between the severity of arterial lesions (arteriosclerosis) and arterial hypertension; moreover, arteriosclerosis and hypertension can be observed separately. Probably functional disturbances provoke structural lesions at a later date. Nothing so far known opposes the theory that human hypertension can be produced by a humoral agent such as hypertensin, but the presence of this substance in human blood has not been demonstrated (see Chap. 63).

Local vascular spasms. *Intermittent claudication due to arterial spasm.* Intermittent functional claudication is a characteristic condition resulting from a local and persistent vasoconstriction (angiospasm). In typical cases, owing to the constriction of the main artery of one or both legs, the circulation is sufficient only at rest or in moderate exercise of short duration, but becomes insufficient in heavy or prolonged exercise. The patient suffers intense pain and difficulty in moving the limb when he makes a heavy or prolonged effort. The pain disappears with rest and walking can be continued, but claudication reappears as soon as the effort is renewed and sustained for a certain time. Incomplete arterial occlusion due to degenerative processes of the arterial walls (thrombosis, embolism) produces the same syndrome. Differential diagnosis between functional claudication due to arterial spasm and that due to an organic arterial lesion is very difficult in some cases.

Many acute and paroxysmal visceral pains have been attributed to arterial spasm.

Raynaud's disease. In this disease there is symmetrical arterial spasm in the fingers of both hands. The fingers become first cold and livid and then cyanotic; there is intense pain as long as the spasm lasts. The angiospastic crises are sometimes so intense and prolonged that trophic lesions are produced, and gangrene of the fingers may occur.

BIBLIOGRAPHY

BRAUN MENÉNDEZ, E., J. C. FASCIOLO, L. F. LELOIR, J. M. MUÑOZ, and A. C. TAQUINI, "Hipertensión Arterial Nefrógena," El Ateneo, Buenos Aires, 1943 (English translation, "Renal Hypertension," Charles C Thomas, Springfield, Ill., 1946).

BURN, J. H., *Physiol. Rev.*, **18**, 137, 1938 (sympathetic vasodilators).

FASCIOLO, J. C., "Hipertensión Arterial Nefrógena," El Ateneo, Buenos Aires, 1939.

GADDUM, J. H., "Gefässerweiterndestoffe," Thieme, Leipzig, 1936.

GAGER, L. T., "Hypertension," Williams & Wilkins, Baltimore, 1930.

GOLDBLATT, H., *Harvey Lect.*, **33**, 237, 1937 (hypertension of renal origin).

GOLDRING, W., R. J. BING, E. CRUZ COKE, *et al.*, "Experimental Hypertension," special publications of the New York Academy of Sciences, vol. III, June, 1946.

HERING, H. E., "Die Karotissinusreflexe auf Herz und Gefässe," T. Steinkopf, Leipzig, 1927.

HEYMANS, C., "Introduction to the Regulation of Blood Pressure and Heart Rate," Charles C Thomas, Springfield, Ill., 1950.

HEYMANS, C., J. J. BOUCKAERT, and P. REGNIERS, "Le Sinus carotidien et la zone homologue cardio-aortique," Gaston Doin, Paris, 1933.

KOCH, E., "Die reflectorische Selbsteuerung des Kreislaufes," T. Steinkopf, Leipzig, 1932.

MASTER, A. M., L. I. DUBLIN, and H. H. MARKS, The Normal Blood Pressure Range and Its Clinical Implications, *J. A. M. A.*, **143**, 1464, 1950.

McDOWALL, R. J. S., "The Control of the Circulation of the Blood," Longmans, New York, 1938.

PAGE, I. H., and A. C. CORCORAN, "Arterial Hypertension: Its Diagnosis and Treatment," Year Book Publishers, Inc., Chicago, 1945.

PAGE, I. H., and A. C. CORCORAN, "Experimental Renal Hypertension," Charles C Thomas, Springfield, Ill., 1948.

PLESH, J., "Blood Pressure and Its Disorders," 2d ed., Williams & Wilkins, Baltimore, 1947.

RICHARDS, R. L., "The Peripheral Circulation in Health and Disease," E. & S. Livingstone, Edinburgh, 1946.

SCHMIDT, C. F., and J. H. COMROE, *Physiol. Rev.*, **20**, 115, 1940 (vascular chemoreceptors).

ZWEIFACH, W. B., and E. SHORR, "Factors Regulating Blood Pressure," The Josiah Macy, Jr., Foundation, New York, 1949.

CHAPTER 21

The Arterial Pulse

THE ARTERIAL PULSE consists in a sensation of expansion felt, simultaneously with the heartbeat, when an artery is slightly pressed between finger and thumb or when it is lightly pressed through the skin and soft tissues against a resistant surface. This sensation is due to the spread along the arteries of a pressure wave originated by the systolic discharge into the aorta. Slight pressure must be exerted on the artery to counteract the permanent distention of the arterial wall, thus facilitating the perception of changes in internal pressure. If the pressure exerted is too great the artery will be completely occluded and the pulse will be suppressed. Systolic pressure waves traveling down the arteries cause a small expansion of their walls, but this movement is not easily seen or felt if a slight external pressure is not applied.

The systolic pressure wave also provokes elongation of the arteries, and this can be clearly seen if an artery is ligated; at each pulse wave the ligature is pushed forward in the direction in which the wave spreads. This movement can be registered by cutting the artery just below the ligature and connecting it with a recording lever.

In arteries that describe a curve the pressure wave tends to open the arch, sometimes with considerable force. The slight rhythmical extension of the leg coinciding with the heartbeat, observed in subjects with one leg crossed over the other, is due to this expansive force. Rhythmical movements of the head synchronous with the heartbeat have the same origin.

All these movements of the artery caused by the spread of the systolic pressure wave are known as arterial locomotion.

The arterial pulse is one of the aspects of body function most frequently examined by the doctor. The development of more direct and informative methods of exploration of cardiac activity has somewhat diminished its importance; the study of the pulse, nevertheless, continues to be an almost indispensable source of information in routine medical examinations. It is therefore necessary to understand its significance thoroughly in order to appreciate the possibilities and limitations inherent in its study.

Registration of the arterial pulse. A graphic record of the arterial pulse gives a good knowledge of its properties, and palpation of the pulse is rendered more valuable in the light of the information acquired by the study of pulse records. The arterial pulse can be recorded by dissecting the artery and inserting in it a cannula connected to a recording manometer, or else by amplifying and recording the oscillations of the arterial walls caused by pressure changes. The former is a record of arterial blood pressure variations and is commonly used in experimental animals, but only in special circumstances in man. Optical manometers of great efficiency should be used (Wiggers, Hamilton). The second method is the one usually applied in the study of the human pulse.

The record obtained when a calibrated manometer is connected with the artery allows the quantitative estimation of the different pressures occurring throughout the cardiac cycle, but when oscillations of the arterial walls are recorded, only a qualitative appreciation of pressure changes can be made. In this latter case, if pressure values are desired, they must be obtained separately by one of the sphygmomanometric methods described in Chap. 19, and only the pressures at the peak (systolic) and the beginning (diastolic) of the pulse wave are measured.

Pulse-wave records obtained by connecting the different arteries with a manometer could

have been described when considering the variations in arterial blood pressure, but this description will be given in this chapter because pressure changes are the fundamental cause of the pulse.

very short time most of the energy developed by ventricular contraction is employed in overcoming the inertia of the column of blood and in distending the arterial walls, and consequently a very small amount is transformed into

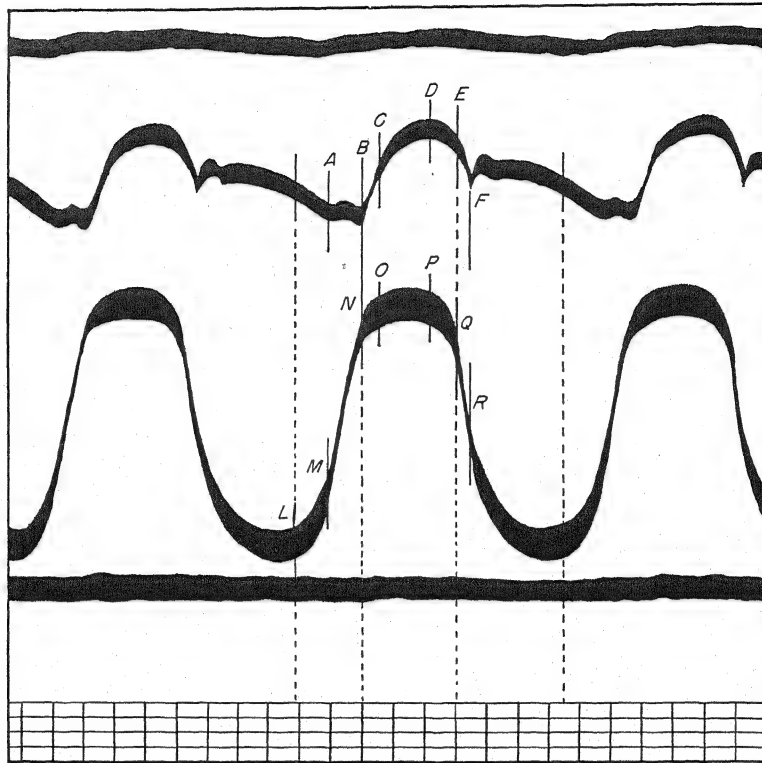



FIG. 105. Aortic and left intraventricular pressure curves. Upper curve (aortic pressure): *A-B*, preliminary wave; *B-C*, minimum ejection subphase; *C-D*, maximum ejection subphase; *D-E*, reduced ejection subphase; *E-F*, aortic incisura. Letters on the lower (ventricular pressure) curve, the same as in Fig. 35. Time in 0.04 sec.

The aortic pressure curve. Figure 105 reproduces a record of aortic (upper tracing) and left intraventricular (lower tracing) pressures, simultaneously taken with Wiggers' optical manometers inserted respectively in the root of the aorta just beyond the semilunar valves and in the left ventricle. Systolic events begin in the aorta with a small wave *A-B*, which coincides with the end of ventricular isometric contraction. This wave and another, which in some cases precedes it and coincides with auricular systole, are Frank's preliminary waves.

Important changes in aortic pressure begin with the opening of the semilunar valves (*N*) and the commencement of the ejection phase. When ample communication between the ventricle and aorta is first established, during a kinetic energy moving the blood forward. This is the subphase of minimum ejection (*B-C*). Then there is a change in the pressure gradient because the systolic impulse is balanced by the runoff toward the periphery. This is the subphase of maximum ejection (*C-D*). When the loss of pressure through displacement of blood toward the capillaries predominates over the pressure developed by the ventricle, the record marks a descending curve of slow gradient (*D-E*). This is the reduced ejection subphase.

At *E* the pressure begins to fall rapidly, and in the records the curve appears as if it had been suddenly cut; this has been called the aortic incisura by Frank. The explanation of this phenomenon usually given is that at *Q* (in the ventricular record) the majority or all of the

ventricular fibers have completed their contraction and the whole ventricle, en masse, begins to relax. The drop in pressure, which continues in the ventricle, is suddenly interrupted in the aorta by closure of the semilunar valves (*F*). This produces a series of sharp oscillations in the arterial pressure, which are marked by vibrations known as postincisural vibrations. The pressure and tension of the arterial walls then fall gradually because of the runoff into the capillaries. The record descends at a slow gradient until at the next systole the cycle begins again.

These pressure changes in the aorta, transmitted in a wavelike manner along the arteries, are the fundamental cause of the pulse. As they travel to the periphery they are increasingly modified. The arterial tubes, because of the elasticity and distensibility of their walls, are better adapted for an efficient utilization of cardiac energy than for the faithful transmission of small pressure oscillations. The fundamental pressure pulse is also modified by the movement of the blood and by reflected waves arising in the periphery of the arterial system.

The carotid pressure curve. Figure 106 reproduces a typical record of pressure changes in the carotid artery of the dog, obtained by inserting a Wiggers manometer in the right carotid, near its origin in the aorta. Curves obtained at this level usually exhibit a small, brief oscillation (*A-B*) corresponding to the preliminary wave seen in the aortic pressure pulse during isometric contraction of the ventricle.

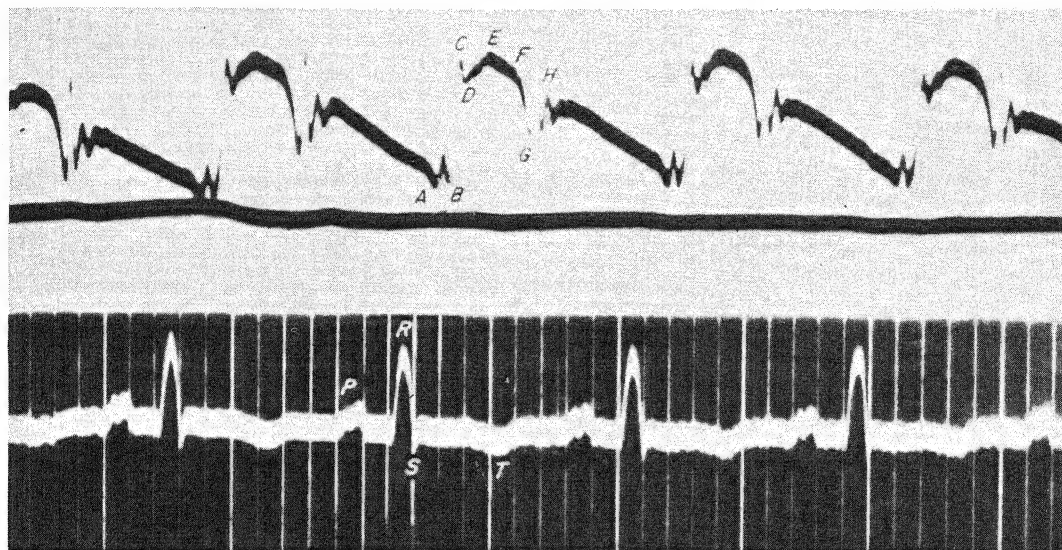



Fig. 106. Carotid pressure curve in the dog, registered with Wiggers' universal manometer (upper curve); simultaneous ECG (lower curve). *A-B*, preliminary wave; *B-C-D*, primary peak; *B-E*, minimum and maximum ejection subphases; *E-F*, reduced ejection subphase; *F-G*, incisura; *H*, postincisural vibrations. Time in 0.2 sec.

After the opening of the aortic semilunar valves the pressure rises in an almost vertical straight line, which is suddenly interrupted by a short fall (*B-C-D*), thus forming a sharp peak. This oscillation, which is not seen in the aortic pulse, is due to the inadequacy of the arteries for the faithful transmission of pressure changes. The curve is the same as that which would be obtained if a cannula were placed in the aorta and connected with the manometer by an elastic tube with the same properties as the artery. The efficiency of the manometer is decreased because the distensible tube diminishes the volume-elasticity coefficient and increases the effective mass of the system. The rapid rise in pressure and the change in gradient of the aortic pulse are therefore not faithfully recorded; they are replaced by the sharp peak just mentioned, caused by the vibration of the liquid mass in the arteriomanometric system. This peak will become more marked the quicker the rise in pressure and the greater the distensibility of the artery. It portrays the natural frequency of vibration of the arteriomanometric system. Usually there is only one

peak, but several vibrations can be observed if the blood pressure is low or when the blood is ejected from the ventricle with unusual force.

A period of rising pressure (*D-E*) follows, showing sometimes small irregularities of the same nature as the primary peak, though less prominent. After reaching a maximum, the pressure falls slowly (*E-F*) until, with the beginning of ventricular relaxation, it falls sharply (*F-G*) as in the aortic incisura. The fall in pressure is suddenly stopped by the closure of the aortic semilunar valves. A kind of rebound causes one or more postincisural vibrations (*H*) in the pressure curve, which continues to descend until the following systole again causes it to rise.

Fig. 107. Marey's direct-transmission sphygmograph.

Registration of the human pulse. Sphygmography and sphygmograms. Sphygmography is the technique used in registration of the pulse by means of instruments called "sphygmographs," which give records known as "sphygmograms." The majority of the instruments have been designed for registration of the radial pulse.

Sphygmographs have been developed from the instrument first invented by Marey.[1] This consisted in a button applied with a certain pressure over the radial artery by a leather bracelet; it transmitted the expansion of the artery to a lever recording the movements on a smoked surface (Fig. 107). Another model transmitted the movements to a tambour, and the changes in pressure in this tambour were recorded by a second one to which it was joined by a rubber tube (Fig. 108). Direct-transmission sphygmographs, such as Jacquet's, are no longer used. Indirect-transmission sphygmographs have been developed by Frank, Wiggers, and others, using the Frank segment capsule for optical registration.

The carotid and subclavian pulses are registered by adapting to the skin, over the artery, a small glass, wood, or metal funnel, 2 cm. in diameter, connected by a rubber tube with a segment capsule for optical registration. A small lateral opening in the tube allows equilibration of the internal and atmospheric pressures.

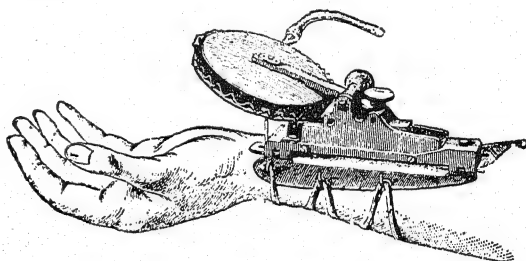



Fig. 108. Marey's indirect (pneumatic) transmission sphygmograph.

The pulse in the arteries of the arms or legs can be registered at different levels by means of a pneumatic cuff, such as those used to measure the blood pressure, joining it to a Marey tambour or a Frank segment capsule through a Chauveau and Marey sphygmoscope (Fig. 109). A sphygmoscope consists of a large

[1] Marey, E. J., "La Méthode graphique," 2d ed., Masson et Cie, Paris, 1885.

glass tube closed by two perforated rubber stoppers. One of the stoppers is covered by a rubber finger stall which is connected by a tube (*a*) to the pneumatic cuff; the other stopper has a small glass tube connected to the Marey tambour or Frank capsule; a lateral opening serves to balance the internal and atmospheric pressures. This apparatus is well suited for recording the pulse in small children, in whom it is difficult to obtain a record of the subclavian or radial pulse.

Fig. 109. Apparatus for optical registration of the arterial pulse of a limb. *M*, manometer; *I*, insufflator; *S*, sphygmoscope; *F*, Frank's segment capsule; *A*, arm cuff; *L*, source of light. (*A. S. Segura, 1937.*) On the right: sphygmoscope. Description in text.

Central, intermediate, and peripheral arterial pulses. The fundamental pressure pulse produced in the aorta by the systolic discharge undergoes considerable changes as it travels down the arteries. Sphygmograms will therefore vary according to the distance from the origin of the aorta at which they are taken. A central, an intermediate, and a peripheral arterial pulse are usually considered. In man the central pulse can be recorded near the origin of the common carotid or subclavian arteries; the intermediate pulse, in the upper segments of the carotid, femoral, and brachial arteries; the peripheral pulse, in the arteries of the distal part of the limbs (radial and pedal arteries).

Terminology. Anacrotic and catacrotic events. The pulse wave in a sphygmogram consists of an ascending limb, a peak or summit, and a descending limb. Oscillations in the ascending limb are called "anacrotic" (Greek *ἀνά*, up, and *κρότος*, stroke), and those in the descending limb "catacrotic" (*κατά*, down). The primary peak in the central pulse is an anacrotic event; the dicrotic wave of the peripheral pulse and the incisura of the central pulse are catacrotic events. The central pulse shows anacrotic and catacrotic oscillations; the peripheral pulse usually has only catacrotic oscillations.

The central pulse in man. The central pulse in man is recorded by placing a funnel in the supraclavicular fossa with sufficient pressure to suppress the pulsation of the jugular vein; a segment capsule connected with the funnel by a rubber tube gives an optical record of the arterial pulse. The features of this pulse are essentially the same as those seen when an optical manometer is inserted into the aorta of a dog. Figure 110 should be compared with Fig. 106. The former is a record of the human subclavian pulse (upper tracing) and the carotid pressure pulse of the dog obtained with a Wiggers optical manometer. Frequently there is no primary peak in the central pulse; in these cases its curve is the same as that of the aortic pressure pulse (Fig. 111).

The following events of the cardiac cycle can be established with great accuracy in a central sphygmogram: (*a*) the duration of the ejection phase (Fig. 110, *B-F*), with the subphases of maximum ejection (*B-E*) and reduced ejection (*E-F*); (*b*) the duration of the protodiastolic

phase (*F-G*); (*c*) the total duration of ventricular diastole. With less accuracy, the preliminary wave (*A-B*) gives the duration of isometric contraction. If, as occurs in some cases, an auricular wave is marked in the sphygmogram, it is also possible to measure the A-V conduction time. The gradient of the upward stroke gives an idea of the velocity of ventricular ejection: the higher the velocity, the steeper the gradient. Finally, in cases of aortic stenosis, vibrations due to the systolic murmur appear in the ascending limb and throughout the greater part of the ejection phase.

The intermediate pulse in man. The sphygmogram of the intermediate pulse obtained by optical registration of the distal part of the carotid, or the brachial artery, diagrammatically represented in Fig. 112 (curve III), differs from the central pulse by the following features: (*a*) preliminary waves are not seen, or are only slightly marked; (*b*) a notch replaces the primary peak; (*c*) the incisura is less sharp; (*d*) postincisural vibrations are damped and tend to fuse into a single wave of longer period. Simultaneous records of the central and intermediate pulses show a greater delay in the beginning of the latter with respect to cardiac events.

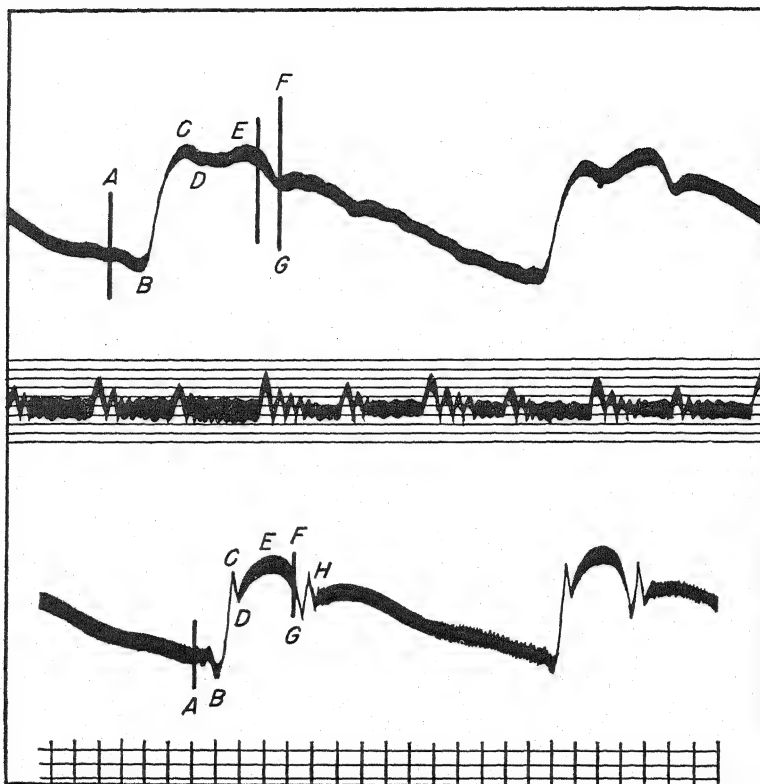



Fig. 110. Central arterial pulse in man (upper curve) and carotid pressure pulse in dog (lower curve). *A-B*, preliminary wave during the isometric contraction; *B-C-D*, primary peak; *B-E*, maximum ejection; *E-F*, reduced ejection; *F-G*, incisura (protodiastole); *H*, oscillations caused by closure of the semilunar valves. Time in 0.2 sec. in the upper record and in 0.04 sec. in the lower.

The peripheral pulse in man. The peripheral sphygmogram, such as is obtained from the radial artery, differs widely from a central pulse record. All minor vibrations are fused in the former into two waves per heart cycle; their contour is smooth and their period long. Curve IV, Fig. 112, is a diagrammatic representation of a radial sphygmogram, optically recorded; Fig. 113 reproduces an optical tracing of the radial pulse simultaneously recorded with the venous pulse. The following features can be easily recognized: preliminary waves have disappeared completely in the peripheral sphygmogram; the wave corresponding to the ejection

phase in the central sphygmogram constitutes the primary or principal wave in the peripheral pulse; it has a smooth contour; the incisura is not marked; closure of the semilunar valves, portrayed as an acute angle in the central pulse, cannot be recognized in the peripheral one. A rounded curve, concave upward, joins without any definite break the primary and secondary (also called dicrotic) waves. The dicrotic wave (Greek δις, two, and κρότος, stroke), so called because when it is pronounced it gives a sensation of a double pulse, is due to the vibrations caused by closure of the semilunar valves, their contour being modified as they are transmitted down the arteries. Sometimes the dicrotic wave is reinforced or its shape modified by interference or resonance phenomena produced by waves reflected from the periphery. It is no longer believed that the dicrotic wave is due mainly to these reflected waves.

FIG. 111. Central pulse in normal human subject. Carotid pulse, heart sounds, ECG, and time in 0.2 sec. The central pulse record is similar to the aortic pressure record.

Circulatory conditions in fever are favorable for the accentuation of the dicrotic wave. Dicrotism is especially developed in typhoid fever, in which there is a clear sensation of a double pulse. (For details on the cause and significance of dicrotism, see Wiggers.[1])

There is a considerable delay of the peripheral pulse with respect to the central pulse (0.05 to 0.08 sec.).

The information given by the radial pulse regarding cardiac function is rather meager. The heart rate and the regularity and relative amplitude of the beats constitute the main data that can be gathered from its study. The central sphygmogram optically recorded is richer in details and can give additional exact points of reference for determining the duration of several phases of the cardiac cycle. In man the only nonsurgical method that permits a detailed analysis of all the phases of the cardiac

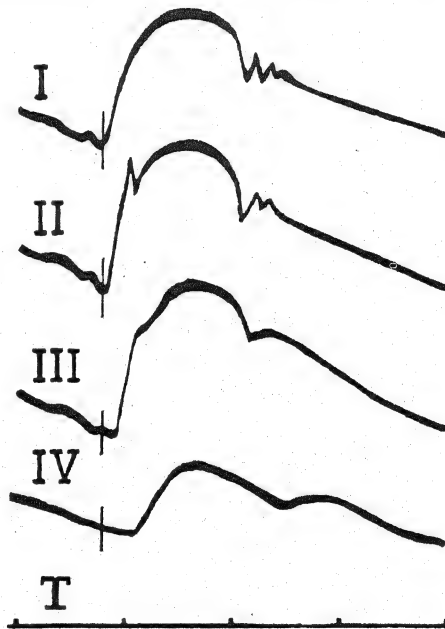



FIG. 112. Diagram showing the changes in the pulse wave as it advances toward the periphery. I, aortic pressure curve; II, pressure curve in an artery close to the heart (central pulse); III, pressure curve in the final part of the common carotid, or in the brachial artery (intermediate pulse); IV, pressure curve in a peripheral artery, such as the radial (peripheral pulse). Time in 0.2 sec.

[1] WIGGERS, C. J., "Physiology in Health and Disease," 4th ed., Lea & Febiger, Philadelphia, 1944, p. 633.

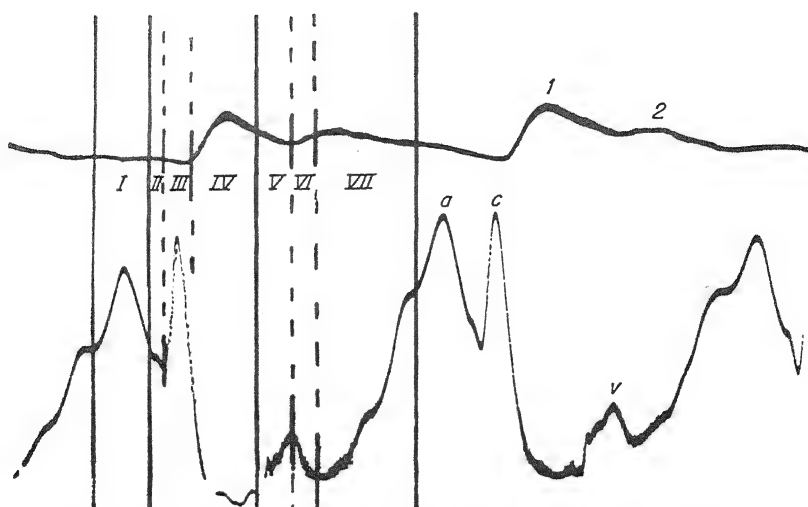



FIG. 113. Radial pulse (upper curve) and phlebogram (lower curve). 1, primary wave; 2, dicrotic wave; I, auricular systole or presystole; II, isometric contraction (approximately); III, delay of arterial pulse; III and IV, ejection phase; V, isometric relaxation; VI, rapid filling of ventricle; VII, diastasis.

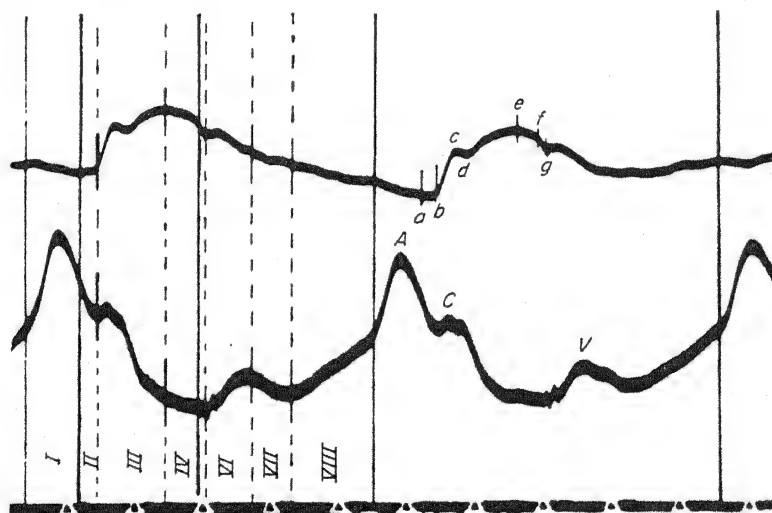



FIG. 114. Phases of the cardiac cycle in man. Upper tracing: central arterial (carotid) pulse. Lower tracing: venous pulse. I, auricular contraction (presystole); II, isometric contraction; III, maximum ejection; IV, reduced ejection; V, protodiastole; VI, isometric relaxation; VII, rapid filling of the ventricle; VIII, diastasis. Time in 0.2 sec.

cycle is the simultaneous registration, by optical methods, of the central arterial and the venous pulses (Fig. 114).

The characteristics of the pulse. The pulse will be considered from a medical point of view, in respect to both the sequence of beats and the characteristics of the waves. Palpation of the radial pulse will be compared with the results of graphic registration, which gives more precise details and a permanent record.

The general characteristics of the pulse are: (*a*) amplitude; (*b*) equality; (*c*) regularity; (*d*) tension; (*e*) frequency.

Amplitude is due to expansion of the artery and is marked on the records by the height of the pulse wave. There is an ample pulse (pulsus magnus) if, on palpation, there is a clear sensation of the fingers being lifted, and a high pulse wave is recorded in the sphygmogram. Apart from factors due to exploratory methods (the degree of pressure exerted on the artery, etc.), amplitude varies as the pulse pressure in the

artery. If the amplitude is small the pulse is known as a small pulse (pulsus parvus).

Equality in the pulse refers to the relative amplitude of the different beats. When the pulse is normal all the pulsations have the same amplitude. An alternating pulse consists in the regular alternation of a large and a small pulsation; this is a typical instance of an unequal pulse. The so-called "paradoxical" pulse (pulsus paradoxus) is another example: the pulse wanes during inspiration, because the blood pressure falls owing to the fall in intrathoracic pressure (see "Respiratory fluctuations" in Chap. 19, Circulation in the Arteries).

Regularity in the pulse refers to the relative duration of the intervals between the pulse waves. The pulse is regular or irregular according to whether the beats are separated by intervals of equal or unequal duration. Premature beats, auricular fibrillation, incomplete heart block, and respiration are the main causes of irregularity in the pulse. The term "rhythm" originally referred to the regularity of the pulse, but it has acquired a wider significance, including three properties of the pulse: equality, regularity, and frequency. If a pulse is equal, regular, and of normal frequency it is a normal rhythmic pulse. If, on the contrary, any or all of these three properties are abnormal there is arrhythmia. Complete arrhythmia, such as occurs in auricular fibrillation, involves abnormality in all three properties.

Tension or hardness of the pulse is measured by the force necessary to suppress the pulse wave and is therefore a qualitative estimation of systolic pressure. It is of little value, because it is entirely subjective, but it can give a general idea of the blood pressure. Sphygmomanometric determinations should always be made. Hypertension causes the pulse to harden (pulsus durus) and hypotension softens it (pulsus mollis). Arteriosclerosis hardens the arteries and can give a false impression of the tension or hardness of the pulse; it should always be borne in mind, especially in elderly subjects.

Frequency of the pulse is given by the number of pulsations per minute; therefore it depends fundamentally on the heart rate. The normal frequency in adults, after a few minutes' rest, is 75 to 85 beats per minute when lying down. Increased frequency is called tachycardia (pulsus frequens); if the rate is 60 or less per minute there is bradycardia (pulsus rarus). Sometimes the pulse rate is lower than the heart rate, because some of the systoles do not eject blood into the arteries; in this case there is bradysphygmia. The heart rate, therefore, cannot be determined with certainty by counting the pulse.

The characteristics of the individual pulse wave are of special interest in some cases, as those of dicrotism already mentioned (fevers, typhoid fever, etc.). In cases of arterial hypertension, and especially in cases of aortic insufficiency, the artery expands and retracts very rapidly; sphygmograms show steep gradients in the ascending and descending limbs of the pulse wave (pulsus celer or Corrigan pulse). In aortic stenosis, on the contrary, the artery expands slowly and there is a low gradient in both limbs of the pulse curve (pulsus tardus).

Velocity of transmission of the pulse wave. The pulse does not begin simultaneously in all the arteries; there is an appreciable interval between the central and peripheral pulses. The speed of transmission can be measured by simultaneously recording the pulse in two distant points of the same artery, or in an artery and one of its branches. If the interval between the waves in the two records and the distance between the points at which they have been registered are accurately measured, the velocity of transmission of the pulse wave can be determined. The figure usually found is 7 to 9 m. per sec.

The average velocity of blood flow in the arteries is 50 cm. per sec., but the pulse is transmitted at much greater speed; therefore the pulse is not caused by the arrival of the blood ejected by ventricular systole to the point where the pulse is registered. A comparison may be helpful: if an object is floating downstream on a river, the movement of this object with respect to the river banks shows the velocity of flow. If now a stone is thrown into the river, upstream with respect to the floating object, the ripples caused by the stone will overtake and pass the floating object, because the speed of these pressure waves is greater than the velocity of flow.

Several factors condition the velocity of transmission of pulse waves. Moëns[1] correlated these factors and expressed the results mathematically. Lambrosi, in a recent study of this question,[2]

[1] Moëns, A. I., "Die Pulskurve," Brill, Leiden, Netherlands, 1878.

[2] Lambrosi, P., *Helvet. physiol. et pharmacol. acta.* **8**, 209, 1950.

concludes that a more accurate expression is given by what he calls the Moëns-Kortweg equation:

$$V = \frac{1}{\sqrt{\frac{d}{E} + \frac{2dR}{a_1E_1}}}$$

in which V is the velocity of the pulse wave, d is the specific weight of the fluid, E is the module of elasticity of the whole tube, a_1 is the thickness of the wall of the tube, and R is the radius of the cross section of the lumen of the tube.

The delay in the transmission of the pulse wave. An interval elapses between the beginning of ventricular systole (as registered by the apex beat or the first heart sound) and the beginning of the pulse wave in the explored artery. This delay is too small to be measured otherwise than by accurate graphic registration. The greater the distance from the heart, the longer will be the delay.

Part of this delay is common to all the arteries, including the aorta. Before the blood can be ejected into the arteries the ventricle must raise the pressure above the diastolic arterial pressure, *i.e.*, there is the isometric contraction phase. The rest of the delay is due to the time the pulse takes to spread down to the point explored; it varies for each artery and increases with the distance from the heart. The pulse at the origin of the aorta shows only the first or essential delay. Other arteries add their own to this delay; it is 0.01 to 0.02 sec. in the common carotid and subclavian arteries, and 0.06 to 0.08 sec. in the radial artery (Fig. 112).

BIBLIOGRAPHY

Bramwell, C., *Lancet*, 2, 239, 301, and 336, 1937 (arterial pulse).

Frey, W., in Bethe's "Handbuch der normalen und pathologischen Physiologie," 7-2, 1223, Springer, Berlin, 1927 (arterial pulse).

Marey, J., "La Méthode graphique." 2d ed., Masson et Cie, Paris, 1885.

Matthes, K., "Kreislaufuntersuchungen am Menschen mit fortlaufend registrierenden Methoden," Thieme, Stuttgart, 1951.

Orías, O., "Registro e Interpretación de la Actividad Cardíaca," 3d ed., El Ateneo, Buenos Aires, 1939.

Weber, A., "Die Elektrokardiographie und andere graphische Methoden in der Kreislaufdiagnostik," 2d ed., Springer, Berlin, 1935.

Wiggers, C. J., "Modern Aspects of Circulation in Health and Disease," Lea & Febiger, Philadelphia. 1923.

CHAPTER 22

Functions of the Capillaries

THE CAPILLARIES ARE the blood vessels with the thinnest walls and the smallest diameter. "Capillary vessel" literally means a vessel as thin as a hair, but the capillaries are much finer than hairs. They divide and anastomose to form a net, which contains in its meshes the cells of the tissues and organs. Blood reaches the capillaries from the arteries, which divide and subdivide, the walls becoming thinner, until without any sharp transition the small arterioles are converted into capillaries. The capillaries converge and form vessels of increasing diameter which, also without transition, are gradually surrounded by muscle fibers until a complete coat, with little elastic tissue, is formed; these vessels are the small veins.

The primary genetic unit of vascular tissue is the capillary. The first vessels to appear in the embryo (Wolff's islets, opaque area) are capillaries formed from mesenchymatous cells. In the early stages of development there are no other vessels. Primitive erythrocytes are also formed by the proliferation of the cells in the capillary walls. After birth, hemopoietic activity is limited to the sinusoid capillaries of bone marrow (see Chap. 5). All the vessels, even the largest, *i.e.*, the aorta, have developed from capillaries; and the heart, although a complicated organ, is at first a capillary tube. Capillaries have therefore a potential capacity to develop into blood vessels of any kind.

The proliferative activity of the capillaries is considerable in the embryo and throughout the period of growth. This activity is responsible for the formation of new vascular territories in the growing tissue. It is never completely lost, and in adult life new capillaries and blood vessels can be formed in response to adequate stimuli. Capillary proliferation is evident in the process of healing wounds, in the organization of clots, and when tissues develop in certain functional conditions, *e.g.*, the uterus and mammary glands during pregnancy. The capacity to form more complicated blood vessels also persists in the adult. When there is gradual arterial occlusion, a new arterial system is developed from the existing arterioles and capillaries. Collateral circulation developed in response to venous occlusion is formed in the same way.

The main object of the circulatory function is the continuous conveyance of blood to and from the capillaries. In these vessels exchanges take place between the blood and the tissue fluids which are the immediate environment of the cells. The chemical composition and physical properties of these fluids, *i.e.*, the conditions necessary for the life of the cells, are dependent on these exchanges. The capillaries are therefore of fundamental functional importance.

Morphology. There is good evidence for the existence of three structural components of the capillary wall: (*a*) the endothelium, made up of pavement-like cells joined by an intercellular cement; (*b*) an endocapillary lining, which is noncellular and is possibly derived by absorption from the circulating blood proteins; (*c*) a pericapillary sheath, with characteristics common to the surrounding connective-tissue matrix, which serves as an external supporting layer.[1]

The endothelium is, undoubtedly, the most important part of the capillary wall. Substances passing into or out of the capillaries probably do so through the intercellular cement, but this cement is produced by the endothelial cells and its physiologic properties are maintained by them. The physical properties of the cement are conditioned by the pH and the calcium content of the surrounding fluid. Normally intercellular

[1] CHAMBERS, R., and B. W. ZWEIFACH, *Physiol. Rev.*, 27, 436, 1947.

cement is being continually renewed. Leukocytes can penetrate through the intercellular cement and migrate out of the capillaries (diapedesis). The breach thus made is rapidly filled, and the integrity of the capillary wall is completely restored.

FIG. 115. Apparatus for direct observation of the capillaries of the skin bordering the nail.

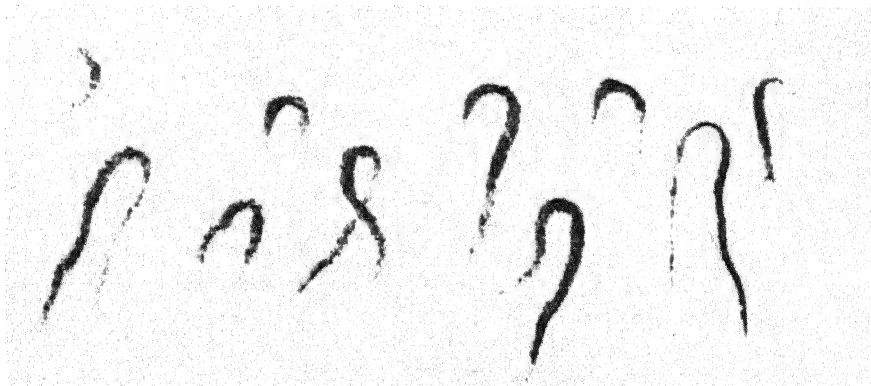



FIG. 116. Capillaries of the skin bordering the nail. The loops are a deep red. Careful observation reveals the movements of the red cells (not evident in picture) and the occasional closure and opening of capillaries.

The average length of a capillary is 0.5 mm. (0.2 to 0.8 mm.), and the diameter varies from 5 to 20 μ. The erythrocytes must therefore change in shape to pass through the narrowest capillaries. An individual capillary is a minute structure, but the whole capillary system is of enormous size. The capillaries of a man weighing 50 kg., if joined end to end, would extend to a length of 100,000 km. (several times around the earth), and their total surface would cover 6,300 sq. m. (about 1½ acres).

Capillaroscopy. Microscopic observation of the capillaries *in vivo* is a common laboratory demonstration, very easy to perform in more or less transparent membranes or tissues such as the mesentery or the lung of the frog.

Human capillaries can be readily observed under the microscope in the skin around the root of the nail, if transparency of the outer layers is increased by treating it with glycerine or mineral oil. A powerful light is reflected on the surface of the skin (Fig. 115). Still and moving pictures of the capillaries can be obtained with a similar arrangement (Fig. 116).

Terminal circulatory units. The shape and distribution of capillaries vary in different organs and tissues. Mostly, however, they form part of structural and functional units. In tissues in which the circulation does not have to respond to complex functional adaptations, the capillaries are not organized. In others, in which there are more versatile functional conditions, *e.g.*, the intestinal tract and the muscles, the distribution

and regulatory mechanisms are more complicated. When the tissue is resting, blood flows only along certain channels; when activity increases, it flows through the whole capillary network.

thousandth of the velocity in the aorta. At this speed an erythrocyte takes about one second to move along the whole length of the capillary.

The blood pressure in the capillaries cannot be easily determined. The most accurate method

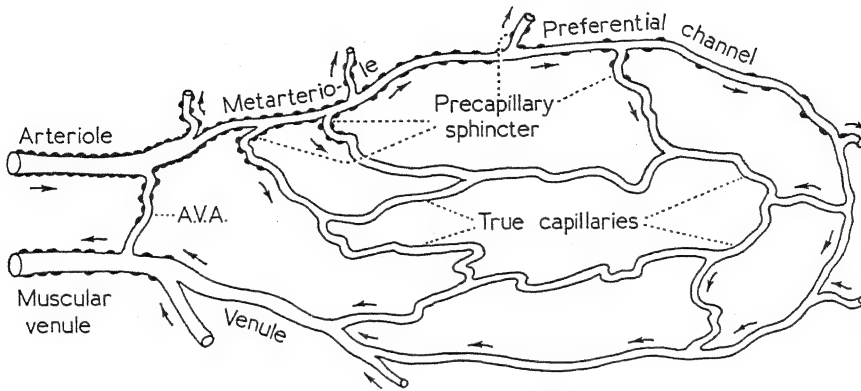



Fig. 117. Schematic representation of the structural pattern of the capillary bed. The distribution of smooth muscle is indicated in the vessel wall A. V. A., arteriovenous anastomosis. (*Zweifach, B. W., in "Factors Regulating Blood Pressure," Trans. Third Confer., Josiah Macy, Jr., Foundation, New York, 1949.*)

The capillaries, anastomosing with each other and connecting with the artery in which they arise and the veins where they end, form a terminal circulatory system[1] (Fig. 117). The arteriovenous channels of preferential circulation act as thoroughfares for blood flow. The metarteriole and the precapillaries at the proximal end of the terminal unit have muscle fibers which intermittently open the precapillary sphincteric offshoots. This mechanism regulates the circulation through the tissues. The true capillaries which follow have no muscle fibers and by multiple anastomoses form a complex network. This network is drained by postcapillaries into the distal portions of the preferential channels. Arteriovenous anastomoses arising above the metarteriole and ending in the venule provide a short circuit which gives even more functional plasticity to the system.

Velocity of flow and blood pressure in the capillaries. The velocity of flow in the capillaries can be determined directly by microscopic observation. The time taken by a red cell to travel a known distance is measured, and the speed is then calculated. The usual velocities are of 0.5 to 0.8 mm. per sec., about one-

[1] Chambers, R., and B. W. Zweifach, in "Factors Regulating Blood Pressure," Transactions of Third Conference, Josiah Macy, Jr., Foundation, New York, 1949.

is to introduce by micromanipulation a microcannula into the capillary, connecting it with a suitable manometer. This technique, developed by Rehberg and Carrier, has been applied in man by Landis.[1] A transparent microcannula filled with saline was introduced into a capillary in the skin. The cannula was connected to a manometer and an insufflator. When the pressure in the system was below the pressure in the capillary, blood flowed out to the cannula, and when the pressure was raised above that in the capillary, saline was injected. By equilibrating the pressures it was possible to measure the exact capillary pressure. The following average values were found in the capillaries of the skin over the nail bed in normal subjects lying down, and with the hand at the level of the manubrium of the sternum (point of reference for the level of the heart): arterial side, 32 mm. Hg; curve of loop, 20 mm. Hg; venous side, 12 mm. Hg. The pressure therefore falls considerably in the short distance between the artery and vein. There is a pulse pressure of 5 to 10 mm. in the arterial end, *i.e.*, capillary pressure when measured in systole is 5 to 10 mm. higher than when measured in diastole. At the venous end there is no appreciable difference between the pressure measured during diastole and systole.

If the finger is lifted above the level of the

[1] Landis, E. M., *Heart*, **15**, 209, 1930.

manubrium, the pressure rises slightly in the arterial side and is not changed in the venous side. If the finger is placed below the level of the heart, the pressure rises in the arterial and venous sides because of the hydrostatic effect. If venous circulation is obstructed, capillary pressure rises up to 10 mm. Hg above the venous pressure. Arterial vasodilatation causes a marked rise in pressure through the entire length of the capillary (from 40 to 60 mm. Hg in the arterial end, and from 10 to 40 mm. Hg in the venous end); pulsation also becomes evident. The increase in pressure caused by venous obstruction is due to the damming effect of the obstacle; that caused by arterial dilatation is due to lowering of peripheral resistance so that a larger proportion of the arterial pressure is transmitted to the capillaries.

Little is known about the pressure in the capillaries of the deeply situated vascular territories. In the glomerular capillaries of the kidney, owing to the proximity of the aorta and to the rapid branching of the renal artery, the blood pressure is much higher than in the skin capillaries; in mammals it has been calculated as being from 70 to 90 mm. Hg. This high capillary pressure is important in the process of glomerular filtration (see Chap. 63).

Capillary pressure is one of the main factors in the exchange of substances between the blood and tissue fluids through the capillary endothelium.

CHANGES IN CAPILLARY DIAMETER

The capillaries in the body are not all open and functioning at the same time. If they were, all the blood would be poured into the capillaries and would stagnate there; the circulation would stop. The number of capillaries open in any territory at a given moment is dependent on the activity of the tissues. If the capillaries of a resting muscle are made visible by injecting India ink, very few of them are seen, and these are of small diameter. When the experiment is repeated in a muscle stimulated by a tetanizing current, the number of visible capillaries increases tenfold, or even more, and their diameters are considerably larger (Krogh) (Fig. 118). The increase of blood flowing through the muscles is compensated by simultaneous vasoconstriction in other territories.

Fig. 118. Diagram of capillaries in the muscles of a guinea pig injected with India ink. From left to right: resting muscle, 200 open capillaries per square millimeter; after moderate exercise, 700 open capillaries per square millimeter; after intense exercise, 2,500 open capillaries per square millimeter. (*After Krogh.*)

Many factors cause the opening and closure of the capillaries.

When the capillaries of an organ dilate, the organ increases its volume but there is no significant change in arterial blood pressure, because peripheral resistance is caused mainly by arteriolar constriction, and cannot therefore be modified to any extent by a dilatation occurring downstream with respect to the arterioles.

Nervous factors. In the frog, stimulation of the sympathetic nerves causes capillary constriction, even when the circulation has been stopped by extirpation of the heart. The capillary constriction is therefore an active phenomenon. Stimulation of the cervical sympathetic in the rabbit produces constriction of the capillaries in the ear of the same side, even when the circulation has been arrested by occlusion of the artery. On the other hand, centrifugal stimulation of the dorsal spinal roots in the frog causes capillary dilatation in the

corresponding territory, even when the arteries have already been maximally dilated by an injection of acetylcholine.[1] In this case dilatation cannot have been a passive result of an increase in capillary blood pressure. The effects of nerve stimulation (sympathetic nerves and dorsal roots) are therefore in the same direction in the capillaries as in the arteries.

Humoral factors. Local capillary dilatation occurring in the course of activity may be caused by substances resulting from the increased metabolism of the tissue. Sometimes they act by modifying the local acid-base equilibrium, *i.e.*, CO_2 and other metabolites lower the pH. Other substances act through a different mechanism. Thus histamine, which is a powerful capillary dilator, is released whenever cells are damaged. Several of the responses observed in the skin capillaries have been attributed to the release of histamine or histaminelike substances (Lewis's H substance).

Several hormones play a part in the control of capillary diameter. Adrenaline and pituitrin exert a marked constrictor effect.

Response of the cutaneous capillaries. The color of the skin. The color of the skin is fundamentally due to melanin, a black pigment found in the pigment cells of the epidermis. A striking characteristic of the white race is the small amount of melanin in its skin; consequently in that race the vessels of the dermis (arterioles, capillaries, and venules) play an important part in giving the skin its color. Through the translucent layers of the epidermis the color of the blood contributes to the color of the skin, according to the number and diameter of the blood vessels and the color of the blood in them. Normally the color is pale pink.

When the concentration of nonoxygenated hemoglobin in the capillary blood rises above 5 gm. per 100 cc., the skin takes on a bluish color, and the condition is known as cyanosis. Methemoglobin (intoxication by nitrites or bismuth subnitrate) and sulfhemoglobin (absorption of sulfur compounds produced by intestinal bacteria) also give the skin a blue color. Accumulation of bilirubin in the blood near or over 20 mg. per liter (normally there is about 6 mg. per liter) confers a yellow tinge to the skin, and the condition is known as jaundice. The intensity of the yellow tinge increases with the concentration of bilirubin in the blood.

[1] Doi, V., *J. Physiol.*, 54, 227, 1920.

Changes in diameter of cutaneous blood vessels modify the color and temperature of the skin. The color is influenced principally by the degree of dilatation of the capillaries and by the color of their contents. The temperature is conditioned by the volume flow, which is controlled by the diameter of the arterioles. A pink, warm skin means arteriolar and capillary dilatation. A pale, warm skin is due to dilated arterioles and constricted capillaries. A pale, cold skin means that arterioles and capillaries are constricted. In very cold weather the exposed skin is cold and blue. In this case the arterioles are constricted and the blood flow is slowed down in dilated capillaries; consequently more oxyhemoglobin is changed to nonoxygenated hemoglobin than when the circulation is more active. Venous obstruction causes cyanosis by a similar mechanism; passive capillary dilatation enhances the bluish tinge of the skin.

DERMOGRAPHISM

The white line.[1] If the skin is stroked with a blunt object without exerting too much pressure, a white line appears gradually, exactly limited to the area stimulated. It lasts for a few minutes and then disappears. This is a normal response, more marked in some subjects than others. The skin of the body, which is less pigmented, usually gives a more visible line than that of the limbs. Microscopic observation shows the white line to be due to local constriction of the capillaries. If the arterioles were also constricted, the contour of the line would not be clearly and exactly limited to the area stimulated, because arterioles supply an irregular area of the skin. The response is observed even when the arteries have been previously occluded or the nerves severed.

The triple response.[2] In sensitive subjects, or when a stronger stimulus (heavy stroking) is applied, the response takes place in three stages.

1. *The red line.* After a short latent period, a red line appears, exactly limited to the area stimulated. It gradually becomes cyanotic because oxyhemoglobin in the stagnated capillary blood is reduced more than normally. The red line is due to capillary dilatation and is observed even in denervated areas.
2. *Erythema or flare.* After a further delay of 15

[1] Cotton, T. F., J. G. Slade, and T. Lewis, *Heart*, **6**, 227, 1917.

[2] Lewis, T., *Heart*, **11**, 109, 1924.

to 30 sec., especially in sensitive subjects, around the red line an erythematous area appears. This flare is due to arteriolar dilatation; its borders are irregular, and the color is not uniform in all the flushed skin. It does not become cyanotic, because the arteriolar dilatation maintains an active circulation through the capillaries. The temperature of the flushed skin rises. Flare is not observed in denervated areas, but cocainization of the nerve trunk does not suppress it, because it is due to an axon reflex.

3. *Wheal.* In especially sensitive subjects a wheal develops over the red line. This wheal projects 1 to 2 mm. above the normal skin and reaches a maximum in about 5 min. Whealing is due to the escape of fluid through the capillary walls. The fluid is similar to blood plasma; it has 4 to 5 per cent protein, with traces of fibrinogen, and it clots when shed. This abnormal passage of fluid from the blood to the tissues is caused by an abnormally increased permeability of the capillary wall.

Action of histamine. Lewis demonstrated that intradermal injection of histamine (diluted 1:3,000) provokes the same triple response as heavy stroking of the skin; *i.e.*, red line due to local capillary dilatation, flare caused by arteriolar dilatation, and wheal due to increased exudation of fluid. The time relations are the same in both cases. This fact suggested that the triple response is due to local release of histamine or of a substance similar to histamine called "H substance" by Lewis.

The triple response can be evoked by many agents: tickling, scratching, freezing, and burning of the skin, electric stimulation, local application of caustic and other substances such as atropine, cantharidin, peptone, allergens, etc. In all these cases there is damage to the cells, and the most plausible explanation is that this causes liberation of a substance (histamine or H substance) that provokes the triple response.

Significance of the triple response. Lewis and his associates have carefully examined the vascular reponses of the skin to the application of external stimuli. They constitute a local, and more or less intense, expression of the more general process known as inflammation. The three stages (triple response) are observed in especially sensitive subjects, or when the stimulus is sufficiently strong; but the response can vary in extension and degree. It is not always as typical as the description given here.

The epidermis clearly plays a protective role. The response is less marked where the epidermis is particularly developed, *e.g.*, on the palms of the hands or soles of the feet. If stimulation is repeated a sufficient number of times over the same area, vascular reactions are followed by proliferation of the epidermis with formation of callosities (corns).

CAPILLARY PERMEABILITY

Exchange of substances between the blood and tissue fluids. The blood vessels form a system of completely closed tubes. The walls of these tubes vary in thickness, but however thin they are they can always be recognized. The blood is never in direct contact with the cells nor does it mix with the tissue fluids. Only in the liver and in the spleen are there special capillaries known as sinusoids, the very thin walls of which seem to be missing in parts, so that the hepatic and splenic cells themselves form the walls of the blood vessel.[1] The usual structure, nevertheless, is such that substances brought by the blood to the tissues must pass through the capillary walls to mix with the tissue fluids, and substances eliminated by the cells into the tissue fluids must also pass through the capillary walls to be taken up by the blood. The mechanisms by which substances pass through the capillary walls and the different factors that condition this passage are still insufficiently known. They are essentially processes of filtration[2] and diffusion[3] through the capillary membrane which take place simultaneously. Factors in these processes are the following: (*a*) the capillary membrane; (*b*) the hydrostatic pressures within and without the capillary; (*c*) the physicochemical condition and concentration of the substances in the blood and tissue fluids.

[1] The absence of a wall in the hepatic and splenic sinusoids is not universally admitted.

[2] Filtration is the passage through a membrane of a solvent, and the substances dissolved, caused by a difference in the hydrostatic pressure on each side of the membrane. When only substances in true solution pass and substances in a colloid dispersion are retained, the term used is ultrafiltration.

[3] Diffusion (dialysis) is the passage of a substance through a membrane caused by a difference in the concentration of the substance on each side of the membrane. The substance always passes from the higher to the lower concentration.

The *capillary wall* is functionally complex, and because of this complexity, it has not yet been possible to make a direct analysis of the phenomena taking place in the membrane. The membrane is permeable to water and substances in true solution both from the blood to the tissue fluid and vice versa. Passage of these substances may be interpreted as if it were a process of filtration. The characteristics of the filtrate demonstrate that the "pores" of the membrane are sufficiently small to prevent the passage of protein molecules and, of course, of the blood cells. The white cells pass through the capillary walls and migrate to the tissues by special movements of the cells (diapedesis).

The size of the "pores" of the capillary membrane depends on its nutritive conditions. The permeability of the capillary walls is definitely influenced by anoxia, endocrine secretions, drugs, toxins, etc. Anoxia increases permeability to the extent of allowing the protein molecules to pass out of the capillaries into the tissue fluid. The internal secretion of the adrenal cortex, or an increase in Ca^{++} ion concentration in the blood, onahe contrary, diminishes permeability to water nd certain salts. A very marked increase in permeability is observed in cases of extensive burns, because of the reabsorption of substances from the damaged tissues. The loss of fluid through the capillary wall in these cases is such that the circulating blood volume may diminish considerably; in some cases to such an extent that the life of the patient may be in danger.

Hyaluronic acid in the connective-tissue matrix is also a structural component of the capillary wall, more especially of the pericapillary sustaining sheath. Hyaluronidase, an enzyme which hydrolyzes hyaluronic acid, acts by increasing capillary fragility and facilitating the rupture of these vessels, rather than as a controlling factor in capillary permeability.[1]

Chambers and Zweifach,[2] through ingenious experiments, demonstrated that filtration takes place through the intercellular cement and that proteins that are too large to pass obstruct the passages in the cement. Factors influencing permeability, according to these observers, modify intercellular cement by contributing to its formation or destruction or by changing its consistency.

The hydrostatic pressure necessary for filtration results from the difference between the blood pressure in the capillaries and the pressure of the tissue fluids. Plasma proteins are hydrophil colloids, which, beside the osmotic pressure exerted by their molecules, have an imbibition pressure; the sum of both is the oncotic pressure. The balance between hydrostatic pressures within and without the capillary and the oncotic pressure of the plasma proteins will condition filtration and the direction in which it takes place. Vasomotor acitivity of the metarterioles and the precapillaries exerts considerable influence on the pressure and velocity of flow in the capillary system, and therefore on the exchange of fluid between the capillaries and the surrounding tissues. Thus, when the muscles of the metarterioles and precapillaries are relaxed, the precapillaries are open and the pressure transmitted by the arteries is spread over the whole capillary network. The speed of flow diminishes considerably in the capillaries and in the preferential channels, so there is a tendency to stagnation in the collecting venules. Intravascular pressure rises, and outward filtration is facilitated. On the other hand, when the metarteriole and precapillary muscles are active, the flow of blood into the capillaries is progressively restricted and eventually takes place only along the preferential channels, coursing rapidly to the venules, and pressure conditions become favorable for inward filtration, *i.e.*, passage of fluid from the tissues into the capillaries.

The nature and concentration of the substances in the blood and tissue fluids are of considerable importance in the interchanges through the capillary walls. The part played by the proteins, as already seen, is a major one. If the concentration of plasma proteins is decreased (*e.g.*, by endovenous injection of saline solutions or of solutions of other crystalloids) the oncotic pressure decreases, the effective filtration pressure increases, and more fluid passes out into the tissues. Fluid thus injected into the blood rapidly escapes from the blood vessels; it is useless to endeavor to increase the blood volume by this kind of treatment. To accomplish that objective a hydrophil colloid, incapable of diffusing out of the capillaries, must be used. The best procedure, however, is to transfuse blood, blood

[1] Zweifach, B. W., and R. Chambers, *Ann. New York Acad. Sc.*, **52**, 1047, 1950.

[2] Chambers, R., and B. W. Zweifach, *J. Cell. & Comp. Physiol.*, **15**, 255, 1940.

plasma, or dried blood plasma redissolved in water.

Plasma-protein concentration may diminish either because of insufficient formation, as occurs when there are certain deficiencies in the diet, or because proteins are lost through the kidneys, as in certain cases of renal disease (nephrosis) in which large amounts of serum albumin pass into the urine (albuminuria). The oncotic pressure diminishes in all these cases, and consequently filtration out of the capillaries increases (see "Edema," further on).

Plasma proteins play their part in the process of capillary filtration because (*a*) they are hydrophilic; (*b*) they do not pass through the pores of the capillary membrane; (*c*) the protein content of the tissue fluids is far below that of blood plasma. If protein concentration were the same in tissue fluids as in blood plasma, the oncotic pressure would no longer be a factor in capillary filtration.

Serum albumin is the most important of the plasma proteins in this respect, because it has a smaller molecule and therefore a larger number of molecules per unit weight than the other proteins. The oncotic pressure (imbibition plus osmotic pressure) will be greater the larger the number of molecules in unit volume (molecular concentration).

If hemoglobin were free in the plasma instead of being enclosed in the erythrocytes, owing to its hydrophilia it would exert an oncotic pressure sufficient to annul the hydrostatic (blood) pressure in the capillaries and filtration would not be possible. This is one of the many advantages to the organism of having hemoglobin in the red cells. Hemocyanin, the respiratory pigment of some invertebrates, has a very large molecule (molecular weight 5,000,000) and therefore exerts only a weak oncotic pressure; it can thus be dissolved in the blood plasma without disturbing capillary filtration.

Crystalloids diffuse rapidly through the capillary membrane, so it is difficult to maintain different concentrations on each side of this membrane for any length of time. They have only transitory effects on the movement of fluids from the blood to the tissues and vice versa. Crystalloids pass to and from the tissues by dialysis or are carried in the flow of water.

Consequences of capillary permeability. The formation of lymph. There is a continuous flow of substances from the blood to the tissues and back from the tissues to the blood. When an organ is at rest and the number of open capillaries is small, the mass of substances transported from the blood to the tissues is approximately the same as that from the tissues to the blood. When an organ becomes active the number of open capillaries is greater, and the blood pressure and flow increase in them. When this occurs the outflow from the blood into the tissues is greater than the inflow in the opposite direction. The balance of fluid and substances is drained by another route. If the organ is a gland with external secretion, this may be excreted by the glandular duct. If the tissue is not a gland of this nature, the excess of substances that cannot be reabsorbed into the capillaries is drained by a special system of closed vessels, called the lymphatics, which have no direct communication with the tissue fluids. They join to form vessels of increasing diameter and eventually end in the large veins near the heart.

In the glomeruli of the kidney the capillary filtrate (glomerular filtrate) is collected by a system of tubes (renal tubes), and part of it is excreted outside the body in the form of urine after having undergone important changes by reabsorption of some substances and addition of others (see Chap. 63).

Edema. Abnormal accumulation of fluid in the tissues is known as edema. It frequently occurs in and under the skin where it can be the subject of direct observation. An edematous skin appears swollen; normal dimples, folds, and wrinkles are less conspicuous. When pressure is exerted with the finger, the skin and tissues do not give the sensation of elastic resistance, and as the fluid is displaced a depression or "pit" remains. In some cases edema spreads to many tissues, and the excess fluid is attracted by the force of gravity to the dependent parts where edema is thus more marked. In generalized edema sometimes the internal cavities, pericardium, pleura, and peritoneum are filled with fluid. This condition is known as dropsy or anasarca. The body weight increases in proportion to the amount of fluid accumulated.

Not only water, but also salts, especially sodium chloride, are retained in the organism, because edematous fluid is isotonic with the blood plasma. Reciprocally, retention of salt causes the accumulation of water, so that diets with little or no sodium chloride hinder the formation of edema and facilitate its reabsorption.

Ion-exchange resins[1] have an interesting effect on edema. In the acid form these resins take up Na^+, K^+, Ca^{++} and other inorganic and organic cations. This combination takes place in the intestine of subjects who have ingested them, and not only the cations in the food but also those in the body tissues are combined and eliminated with the resin. In cases of edema the resins are given in order to prevent absorption of sodium (sodium chloride can thus be added to food for its savory properties) and to withdraw sodium (and therefore water) stored in edema. Synthetic resins derived from polystyrene with crossed di-vynil benzene bonds are used. It must be remembered that their action is not specific for any particular cation; they bind all of them, and so they may provoke deficiencies in K^+, Ca^{++}, or Mg^{++} or acidosis.

There are many causes of edema, and in some cases more than one plays a part; in all of them there is an insufficient reabsorption of fluid from the tissues.

Stasis due to venous obstruction or heart failure increases capillary pressure and therefore also the effective filtration pressure. This is not the only cause of edema in such conditions. Anoxia and accumulation of metabolic products due to incomplete oxidation and impaired circulation, by increasing capillary permeability, also contribute to produce edema.

Renal disturbances cause edema through several mechanisms. In cases of nephrosis because of the marked albuminuria, the plasma protein concentration diminishes, and the oncotic pressure decreases; therefore the increased effective filtration pressure drives a greater amount of fluid out of the capillaries. A similar condition can be produced experimentally in animals by plasmapheresis: blood is drawn, and the erythrocytes suspended in saline are reinjected. If this procedure is repeated several times, the plasma protein concentration is significantly lowered and edema results. When the plasma proteins are restored to their normal concentration, edema disappears. In cases of glomerulonephritis, water, salts, and nitrogenous substances are retained; this together with loss of plasma protein by albuminuria produces edema.

Nutritional edema is observed when a qualitatively or quantitatively insufficient diet causes hypoproteinemia. Hypoproteinemia is not, however, always the cause of nutritional edema. In these cases, as in others in which there are cachectic conditions, the pathogenesis of edema is obscure.

Lymphatic obstruction produces edema in the territories normally drained by the obstructed lymphatic vessels. It may be caused by parasites (filariasis, elephantiasis), tumors, etc.

[1] Dock, W., *Am. Heart J.*, **40**, 638, 1950.

Inflammatory edema is due to increased capillary permeability produced by the toxic substances that cause the inflammatory process (bacterial toxins, venoms, caustic substances, etc.) or by substances discharged by the cells when damaged (traumatism, several physical agents such as heat, cold, sunlight, etc.). These substances are histamine or similar to histamine (see "Action of histamine," page 214). The intensity of the process of increased permeability and the area to which it extends vary according to the causative agent. In some cases there is simply an increased transudation of normal fluid; in others the fluid shed has a high protein content, and even fibrinogen may filter through the capillary membrane. All the intermediate conditions are possible between these extremes.

Capillary fragility. The resistance of the capillaries to mechanical force acting on them varies in different subjects and in the same subject in different circumstances. In some cases the slightest trauma, such as moderate pressure on the skin, or the small increase in capillary blood pressure caused by coughing, or a bodily effort, provokes the rupture of the capillaries. Blood seeps into the tissues, a hematoma is formed, and the hemoglobin suffers a series of chemical changes that provoke marked discoloration in the bruised area; finally the blood is reabsorbed.

Capillary resistance to mechanical agents diminishes considerably in certain pathologic conditions, *e.g.*, scurvy. Vitamin C, the lack of which causes scurvy, has therefore been thought to be an important factor in the prevention of capillary fragility. More recently, vitamin P (citrine) and rutine,[1] a flavonol glycoside, have been considered necessary factors for the maintenance of capillary resistance (see Chap. 49).

There are several methods for determining capillary resistance. They are based on two general principles:

1. Capillary pressure is increased by applying an external pressure sufficient to suppress the venous circulation, but not enough to occlude the arteries, so that blood will continue to flow into the capillaries (a sphygmomanometer is used).

[1] Griffith, J. Q., J. F. Couch, and M. A. Lindauer, *Proc. Soc. Exper. Biol. & Med.*, **55**, 228, 1944; Scarborough, H., *Biochem. J.*, **39**, 271, 1945; Shanno, R. L., *Am. J. M. Sc.*, **211**, 539, 1946.

2. The pressure in the tissues surrounding the capillaries is reduced by means of a suction chamber (*e.g.*, cupping).

Tey[1] found a close relationship between the pressures that provoke the rupture of the capillaries and the arterial blood pressure. Capillary fragility is normal when the aspiration pressure must be kept for 5 min. at more than 60 mm. (at least 40 mm.) above the systolic pressure in order to produce moderate hemorrhagic spotting of the skin. It is subnormal when an aspiration pressure of 20 to 40 mm. above the systolic pressure produces this effect, and it is pathologic when hemorrhagic spotting occurs with even lower aspiration pressures. He found that sodium, potassium, and calcium have no effect on capillary fragility, that adrenaline reduces it for a very short time, and that histamine increases it. Again according to Tey, vitamin C diminishes an already increased capillary fragility in some cases but not in others.

[1] Tey, A., "Fragilidad Capilar Normal Humana," Espasa-Calpe, Barcelona and Buenos Aires, 1940.

TRAUMATIC SHOCK

Traumatic shock is characterized by a deep, generalized functional depression following extensive damage to the tissues. It is called "primary shock" when it occurs immediately after the trauma, and "secondary shock" when it appears only after an interval of several hours.

The exact mechanism leading to shock, especially to secondary shock, is still under discussion. When shock has developed, characteristic blood changes and circulatory disturbances are observed. The capillaries play a fundamental part in the mechanism of shock, and for this reason the condition is considered in this chapter.

The blood shows an increase in concentration (hemoconcentration). This causes an increase in the red cell count and volume, in hemoglobin concentration, in the white count, and in the specific gravity of the whole blood. Plasma proteins and nonprotein nitrogen also increase. The condition is one of anhydremia (loss of water from the blood). In cases with profuse hemorrhage blood concentration is sometimes masked by compensatory reabsorption of tissue fluids.

There are considerable disturbances in the circulation. Arterial blood pressure falls to a low level and the volume flow diminishes. It has been definitely proved this is not due to central circulatory insufficiency (heart failure). The fundamental properties of the myocardium are normal, but because the venous return is considerably reduced, the heart does not eject at each systole a sufficient amount of blood to maintain the arterial pressure.

There is peripheral circulatory insufficiency.[1] Arterial hypotension and the diminished volume flow are not due to arterial vasodilatation; on the contrary, the arteries are constricted. Arterial constriction is a compensatory response to hypotension, produced by a reflex starting in the cardioaortic and carotid pressoreceptors. Tachycardia—always observed in shock—has the same origin. Increased sympathicoadrenal activity is thus a compensatory response to hypotension; nevertheless it can increase the severity of shock because it ultimately reduces the mass of circulating blood.[2] Circulatory insufficiency is due to hypovolemia, caused mainly by anhydremia produced by the escape of fluid brought about by an increase in permeability of the capillary walls. Hemorrhage, if it has occurred, accentuates hypovolemia, but even in cases in which there has been no hemorrhage, severe hypovolemia can be observed. In cases without hemorrhage, hypovolemia is caused exclusively by the loss of fluid due to increased capillary permeability. As time passes this disturbance becomes more marked and is finally irreversible.[3]

In experimental shock provoked by hemorrhage in the dog, Shorr, Zweifach, and their associates[4] have distinguished a hyperreactive from a hyporeactive phase. During the former a vasoexcitor material (VEM) appeared in the blood. This substance was demonstrated by injecting it intravenously into rats and observing the sensitizing effect it had on the response to adrenaline by the blood vessels of the meso-appendix. When hemorrhage caused severe hypotension, release of VEM diminished simultaneously with the decrease in renal blood flow, and the hyperreactive phase ended. This phase could not be demonstrated in animals with the kidneys excluded from the circulation by tying the renal vessels. During the hyporeactive phase the liver, and in a lesser degree the spleen and skeletal muscles, released a vasodepressor material (VDM). *In vitro* experiments showed that both substances were formed in response to anoxia. VDM has been identified with ferritin, or its iron-free derivative, apoferritin.

[1] Harrison, T. R., "Failure of the Circulation," Williams & Wilkins, Baltimore, 1935.

[2] Freeman, N. E., *Am. J. Physiol.*, **103**, 185, 1933.

[3] Moon, V. H., "Shock and Related Capillary Phenomena," Oxford, New York, 1938.

[4] Shorr, E., B. W. Zweifach, and R. F. Furchgott, *Ann. New York Acad. Sc.*, **49**, 571, 1948; Mazur, E., and E. Shorr, *J. Biol. Chem.*, **176**, 771, 1948; Shorr, E., B. W. Zweifach, R. F. Furchgott, and S. Baez, *Circulation*, **3**, 42, 1951.

Treatment of shock must be directed in the first place to the restitution of the circulating blood volume with the least possible delay. Intravenous injections of saline or gum acacia are no longer used because they give only transitory and inconstant results. Transfusion of whole blood or plasma, according to the degree of hemoconcentration, must be performed as soon as possible. Shock should be prevented by appropriate measures, and constant vigilance exercised to recognize the initial signs and symptoms (hypotension, progressive tachycardia, hemoconcentration) as soon as they appear. The physician should keep in mind that treatment is very effective for the prevention of shock or in the first stages; later, when the condition has progressed to a certain degree of severity, it becomes irreversible and treatment is no longer effective.

Blood changes and hemodynamic conditions in the principal types of shock. Richards,[1] after careful observation of the blood changes and hemodynamic conditions in different types of shock, has summarized his conclusions as follows:

1. *Secondary traumatic or hemorrhagic shock.* Total blood volume diminished; loss of erythrocytes greater than loss of plasma; venous return diminished; minute volume diminished; arterial blood pressure decreased; intra-auricular pressure decreased; peripheral resistance diminished.
2. *Shock due to burns.* Total blood volume diminished; hemoconcentration; minute volume diminished; arterial blood pressure normal or high; intra-auricular pressure normal or high; peripheral resistance increased. Transfusion of plasma causes a rapid fall of intra-auricular pressure.
3. *Medical shock (pneumonia, sepsis).* Total blood volume normal; minute volume diminished; arterial blood pressure decreased (vasomotor collapse).
4. *Neurogenic or primary shock (syncope).* Peripheral resistance diminished.

[1] Richards, D. W., *Bull. New York Acad. Med.*, **22**, 630, 1946.

Peripheral resistance can be quantitatively gauged, with satisfactory approximation, by dividing the mean arterial pressure by the minute volume. Normal resting subjects have a mean blood pressure of 90 mm. Hg and a minute volume of 5.4 liters; peripheral resistance is therefore equal to 16.

BIBLIOGRAPHY

Blalock, A., "Principles of Surgical Care. Shock and Other Problems," Mosby, St. Louis, 1940.

Blalock, A., *Am. Heart J.*, **23**, 145, 1942.

Cannon, W. B., "Traumatic Shock," Appleton-Century-Crofts, New York, 1923.

Chambers, R., and B. W. Zweifach, Intercellular Cement and Capillary Permeability, *Physiol. Rev.*, **27**, 436, 1947.

Harkins, H. N., *Surgery*, **9**, 231, 447, and 607, 1941 (shock).

Krogh, A., "The Anatomy and Physiology of Capillaries," 2d ed., Yale University Press, New Haven, 1929.

Landis, E. M., *Physiol. Rev.*, **14**, 404, 1934 (capillary pressure and permeability).

Lewis, T., "Blood Vessels of the Human Skin and Their Responses," Shaw and Sons, London, 1927.

Moia, B., "Insuficiencia Circulatoria Periférica. Colapso. Shock," El Ateneo, Buenos Aires, 1942.

Moon, V. H., "Shock: Its Dynamics, Occurrence and Management," Lea & Febiger, Philadelphia, 1942.

Müller, O., "Die Kapillaren des menschlichen Körperoberfläche," Enke, Stuttgart, 1939.

Richards, D. W., *Bull. New York Acad. Med.*, **20**, 363, 1944 (circulation in human shock).

Scudder, J., "Shock: Blood Studies as Guide to Therapy," Lippincott, Philadelphia, 1940.

Shorr, E., B. W. Zweifach, R. F. Furchgott, and S. Baez, Hepatorenal Factors in Circulatory Homeostasis, *Circulation*, **3**, 42, 1951.

Smirk, F. H., Pathogenesis of Edema, *Proc. Roy. Australasian Coll. Physiol.*, **5**, 80, 1950.

Wiggers, C. J., "Physiology of Shock," Commonwealth Fund, Harvard University Press, Cambridge, Mass., 1950.

CHAPTER 23

Circulation in the Veins

AFTER THE BLOOD leaves the capillaries, it flows through the veins. These are converging tubes, arising in all the tissues, and ending in the right auricle by means of two main collecting trunks. This is the general venous system, which corresponds to the aortic system. In the pulmonary system the veins arising in the lung capillaries converge to main trunks ending in the left auricle.

The venous tubes have fibromuscular walls, which are flaccid and collapsible; they increase in diameter as they approach the heart. A prominent and important feature is the existence of valves, placed so that blood can flow only toward the heart. The valves were discovered by Fabritius of Acquapendente[1] in 1574, but their functional significance was not understood until Harvey demonstrated it in 1628.

The veins are formed by a thin wall of fibrous tissue, with few elastic and muscle fibers. The inner aspect of the veins is completely lined by endothelium. The valves are formed mainly by a fold in the endothelium.

Veins are clearly different from arteries, not only in structure but also in other respects. The ratio between the diameter of the vessel and the thickness of its wall is much greater in the veins than in the arteries. Distensibility is different in each type of vessel. An increase in internal pressure rapidly increases the capacity of the veins, but a further increase in pressure rapidly causes maximum distention. Arteries are relatively more distensible than veins. Roy[2] observed that in order to double the capacity of a vein the internal pressure must be raised from 0 to 500 mm. of water. The blood in a dilated vein therefore is not necessarily at high pressure. As the pressure in the large cervical and thoracic veins never reaches 500 mm. of water (the pressure necessary to double their capacity), changes in volume are due to variations in blood mass more than in pressure.

The muscles of the veins are functionally important. If they contract, the diameter and capacity of the veins diminish, and sometimes considerably. These muscles are under nervous control, but humoral, mechanical, and other stimuli acting locally also modify their activity. The large veins near the heart show rhythmic contractions of their walls, synchronous with the heartbeat. In certain animals other veins also contract rhythmically, *e.g.*, those in the wings of the bat.

Venous blood pressure. Direct manometric measurements of venous blood pressure give the most accurate results. The cannula should be placed in the vein without obstructing the blood flow, because the lateral and not the end pressure is important. If the vein is obstructed, blood is dammed up and the pressure rises until it can flow through collateral veins. There is a considerable rise in venous pressure when the principal vein of a large area is closed, *e.g.*, when the principal vein of a limb is tied venous pressure rises up to the level of the mean arterial pressure. To avoid this, T cannulas are used. In the large veins needles can be thrust through the wall. Although the needles must be of sufficiently large bore they must not completely obstruct the vein. A system of tubes connects the cannula or the needle with a water manometer or with an aneroid manometer calibrated in millimeters of water. Burch and Winsor s aneroid flebomanometer[1] is easy to handle and a useful, practical instrument.

[1] Hyeronimus Fabritius Acquapendente (1537–1619), "De venarum ostiolis," 1603, facsimile ed. with notes by K. J. Franklin, Charles C Thomas, Springfield, 1933.

[2] ROY, C. S., *J. Physiol.*, **3**, 136, 1881.

[1] BURCH, G. E., and T. WINSOR, *J. A. M. A.*, **123**, 91 1943.

The vein chosen for the determination of the venous blood pressure should be at approximately the same level as the right auricle. If it lies above this level, the effect of gravity will tend to empty the vein and lower readings than the true ones will be obtained; whereas if it is below, the hydrostatic pressure of the column of blood between the auricular level and the cannula will be added.

Table 18 gives the average venous pressure in several territories of normal subjects; there are considerable individual variations. Venous pressure is not related to age or race; the figures found in women are, however, somewhat higher than those found in men.[1]

Venous pressure can be measured in human beings with no discomfort by resorting to indirect methods based on Von Recklinghausen's method.[2] In this method a cutaneous vein is compressed through the skin by a pneumatic rubber ring, connected to a water manometer and an insufflator. The ring is kept in place by a glass plate through which the exact moment the vein collapses can be noted. The minimum pressure necessary to bring about this collapse is equivalent to the venous pressure. Neither this method nor its modifications give precise results; moreover, they can be applied to only a few venous areas.

An approximate idea of venous pressure can be obtained by very simple procedures. If the subject is lying on his back with his head slightly raised, the jugular vein acts as a manometer. In normal conditions it appears distended up to 1 or 2 mm. above the clavicle. Another method (Gärtner)[3] is to raise passively the hand of a seated subject and observe the exact level at which the veins of the back of the hand collapse. The difference in height between this level and the sternocostal angle (the level of the heart) gives the venous pressure in millimeters of blood (approximately equivalent to millimeters of water) in the veins of the back of the hand. This pressure is normally from 80 to 120 mm.

Venous pressure diminishes progressively from the capillaries to the right auricle. This pressure difference is the principal cause of the

Table 18. Venous Pressure in Normal Subjects

Territory	*Vein*	*Pressure, mm. H_2O*
Head and neck	Angular	166
	Frontal	172
	Ext. jugular	72
Shoulder	Cephalic above the clavicle	101
Arm		
Upper third	Cephalic	83
	Basilic	50
Middle third	Cephalic	94
	Basilic	55
Lower third	Cephalic	99
	Basilic	83
Elbow	Cephalic	97
	Basilic	91
	Med. cubital	103
Forearm		
Upper third	Cephalic	106
	Accessory cephalic	87
	Med. antibrachial	108
	Basilic	107
Middle third	Cephalic	111
	Accessory cephalic	91
	Med. antibrachial	113
	Basilic	110
Lower third	Cephalic	118
	Med. antibrachial	117
Wrist	Cephalic	139
Hand	Dorsal arch	139
Trunk		
Upper	....	120
Lower		140
	Thoracoepigastric	
	At the nipple	96
	At the xiphoid	108
Pubis	....	156
Penis	Dorsal	156
Thigh		
Upper third	Superficial ant.	142
	Saphenous	100
Middle third	Superficial ant.	138
	Saphenous	104
Lower third	Superficial ant.	127
	Saphenous	101
Ankle	Saphenous	126
Foot	Med. marginal	152
	Lat. marginal	145
	Dorsal arch	188

Source: Ochsner, A., R. Colp, and G. E. Burch, *Circulation*, 3, 674, 1951.

[1] Ochsner, A., R. Colp, and G. E. Burch, *Circulation*, **3**, 674, 1951.

[2] Von Recklinghausen, F., *Arch. f. exper. Path. u. Pharmakol.*, **55**, 468, 1906.

[3] Gärtner, G., *München med. Wchnschr.*, **50**, 2038 and 2060, 1903; **51**, 212, 1904.

movement of the blood from the capillaries to the heart. It is normally maintained by the activity of the heart, which is the cause of the pressure in the capillaries, and by the thoracic "negative" pressure. The effective pressure for venous circulation can be maintained, in normal and in pathologic conditions, by pressure changes in the initial (capillary) and final (thoracic) portions of the system. If thoracic pressure is raised by keeping the glottis closed during forced expiration (Valsalva's experiment), this increases the end pressure in the venous system; automatically stasis is produced and the capillary pressure increases, thus restoring the effective pressure needed to keep the venous blood circulating toward the heart. Pulmonary emphysema causes permanently similar conditions. The muscles of the veins can also participate in adjusting the effective pressure.

Posture and exercise have a marked effect on venous pressure. In a series of observations the average venous pressure at the ankle was found to be 11.7, 56.0, and 86.8 mm. H_2O with the subject lying down, sitting, and standing up respectively. Walking caused an appreciable decrease in the venous pressure with respect to its value when the subject was standing still. During the first 3 to 12 paces, venous pressure fell. It then remained stationary at average levels of 22.3, 24.3, and 23.6 mm. H_2O when the subject was walking at 1.7, 2.6, and 3.3 miles/hr. respectively, remaining at this level for the rest of the walk. The level of stabilization was not therefore influenced by the speed of walking. When walking ceased the venous pressure returned to the initial level in ½ min.[1]

Venous pressure gives valuable information to the physician in certain circulatory disturbances, but the influence of variations in intrathoracic pressure and other extracirculatory factors must first be excluded. For example, pulmonary emphysema causes venous hypertension by the mechanism explained above, and mediastinal tumors increase venous pressure by compression and obstruction of the large thoracic veins.

In circulatory failure of central origin, *i.e.*, caused by heart insufficiency, venous pressure is increased. In peripheral circulatory failure (hemorrhage, arterial vasodilatation, capillary hypotonia, shock, etc.) venous pressure falls.

[1] POLLACK, A. A., and E. H. WOOD, *J. Applied Physiol.*, 1, 649, 1949.

Velocity of blood flow in the veins. Velocity of flow in the veins is measured by the same methods as velocity of arterial flow. In experimental animals it has been found to be constant in peripheral veins and to fluctuate in the veins near the heart. Velocity of flow increases from the periphery to the heart, because of the progressive decrease in total cross-sectional area. In the large venous trunks, variations in intrathoracic pressure produced by respiratory movements and the cardiac cycle are responsible for periodic changes in the velocity of flow. The combined effects of all these factors produce changes in volume in the cervical veins, which are easily seen in subjects lying down; this is the venous pulse, which will be examined in detail further on, because of its physiological and medical importance. Fleck[1] measured the velocity of flow in the external jugular vein of dogs by means of a hemotachometer with optic registration. The volume flow as 78.7 cc. per min. during inspiration and 89.2 cc. during expiration, and the respective velocities 215 and 243.5 mm. per sec. In the superior vena cava the volume flow was 395 cc. per min. during expiration and 485.7 cc. during inspiration; and the velocities 183 and 224 mm. per sec. Maximum acceleration and increase in blood flow were observed at the end of inspiration. The maximum reduction in flow and velocity occurred immediately before inspiration. Cardiac activity causes fluctuations of lesser amplitude than the respiratory variations.

The speed of flow in the veins can be measured in man by injecting a radioactive sodium isotope (Na^{24}) into a dorsal metatarsal vein and recording its arrival at the groin by means of a Geiger-Müller counter. The average time in healthy individuals has been found to be 18 ± 0.9 sec., with a range from 4 to 50 sec.[2]

Factors that influence venous blood flow. The principal cause of the movement of the venous blood is the difference in the pressures in the capillaries and in the auricle. This effective pressure is due in the first place to the activity of the heart, which is therefore the efficient cause of the venous return. There are other accessory factors, such as the negative intrathoracic pressure, which also contribute to the venous circulation. Inspiration increases the

[1] FLECK, S., *Ztschr. f. Kreislauff.*, 25, 504, 1934.
[2] WRIGHT, H. P., S. B. OSBORN, and D. G. EDMONDS, *Lancet*, 2, 767, 1948.

"negative" thoracic pressure and expiration decreases it, but in any case this factor is of secondary importance. The heart also exerts a certain sucking effect on the blood of the veins at the beginning of auricular diastole and during the phases of ejection and of rapid filling of the ventricle.

Venous circulation is also influenced by local and general variations in the capacity of the venous system. These variations are due to contraction and relaxation of the venous muscles and to changes in the circulating blood volume. Widespread venous constriction and an increase in blood volume increase the venous return, and vice versa.

Hydrostatic pressure due to gravity acts differently according to the position of the subject and the venous territory considered. In a man standing at ease, gravity is a positive factor in the circulation in nearly all the territory drained by the superior vena cava, and a negative factor in the territory of the inferior vena cava. The negative effect of gravity is antagonized to a certain extent by the venous valves, the vasomotor responses of the venous muscles, and the contraction of skeletal muscles.

The effect of muscular contraction is clearly demonstrated in the following observation made by Atzler and Herbst:[1] if a subject is sitting with his legs hanging loosely, the muscles being relaxed, the volume of the lower part of both legs increases 1.1 per cent in 1 hr.; if he is standing up, the volume increases 3.6 per cent; but if he is walking, the increase is only 1.99 per cent. The increase in volume is due to venous stasis which raises the capillary pressure and causes a greater filtration of fluid from the blood to the tissues. When the muscles contract they press on the veins and empty them; the venous valves direct the flow toward the heart.

Special adaptation reactions take place in abnormal conditions. If the heartbeat weakens or ceases, venous circulation does not stop immediately, even if the arterial pressure falls to zero. This is observed when the heart is stopped by stimulation of the vagus nerve and in ventricular fibrillation. In both cases the heart dilates progressively because its contents continue to increase, and the veins near the heart show considerable engorgement. If an opening is made in the heart or the vena cava, blood flows copiously and more than 50 per cent of the blood volume can be collected in this way, in spite of the nonexistence of the heartbeat. Henderson attributes this flow to the tone of skeletal muscles. It is quite probable, however, that the muscles of the veins, stimulated directly (anoxia) or reflexly (arterial hypotension acting on the cardioaortic and carotid sinus, pressoreceptors) play an important role. Whatever its immediate cause, this mechanism is a safety factor when the heartbeat weakens, because the increase in venous flow and pressure assures the filling of the heart and the increase of initial pressure, which, according to Starling's law, will increase the force of myocardiac contraction. Venous hypertension observed in the first stages of central circulatory insufficiency is due to the same mechanism.

[1] ATZLER, E., and E. HERBST, *Ztschr. f. d. ges. exper. Med.*, **38**, 137, 1923.

THE VENOUS PULSE

Variations in intrathoracic pressure caused by respiratory movements, and changes in intra-auricular pressure in the course of the cardiac cycle, determine fluctuations in the velocity of flow in the large venous trunks near the heart. Changes in volume in the cervical veins thus occur and can be easily seen in recumbent subjects. These changes in volume are the cause of the venous, or jugular, pulse. There is therefore a respiratory venous pulsation and a cardiac venous pulsation. Usually by "venous pulse" is meant the pulsation of cardiac origin, which is of great physiological and medical importance.

Hunter (1794) was the first to observe the venous pulse, which was graphically registered for the first time by Friedereich in 1866. Potain, Marey, Mackenzie, Wenkebach, Ohm, and Wiggers used successively more accurate and sensitive methods of registration and were thus able to establish the nature and significance of the venous pulse in normal and pathologic conditions. Registration of the venous pulse is now commonly used in the interpretation of many aspects of normal and abnormal cardiac function.

The phlebogram. This is the name given to the record of the venous pulse. A small funnel (which is sometimes covered by a very thin rubber membrane) is placed over the external jugular vein in the supraclavicular fossa, exerting slight pressure so as to assure perfect contact between the skin and the entire edge of the funnel. Changes in venous volume are trans-

mitted by a system of rubber tubes to a Marey tambour (registration on a smoked surface) or a Frank segment capsule (optical registration). Optical registration is preferable because records obtained on a smoked surface are deficient in many respects. The subject must be lying down,

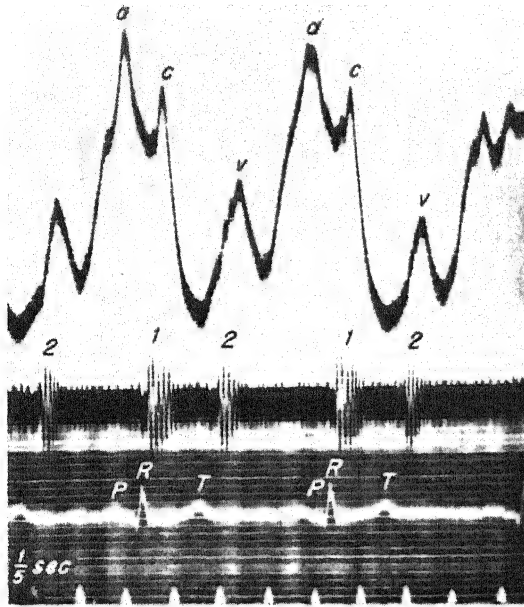



Fig. 119. "Pure" venous pulse. Phlebogram, phonocardiogram, and ECG; simultaneous records obtained from a normal man. The *c* wave is of small amplitude. (*Caeiro, 1942.*)

to avoid the effect of gravity which tends to empty the vein and diminish the amplitude of the venous pulse. The head should be resting comfortably, so as to obtain complete muscular relaxation, and should not be turned to the opposite side.

At each cardiac cycle the normal phlebogram constantly shows three main waves and two depressions[1] (Figs. 28, 39, 40, 45, 46, 113, 114, 119, 120). These features have been given different names. Mackenzie's terminology has been generally adopted; he named the main waves *a*, *c*, and *v*, in the order of their appearance in the cardiac cycle, because the first one was attributed to the auricle, the second to the carotid, and the third to the ventricle. The depressions were called *x* and *y*. In a long cycle sometimes a fourth wave (*h*) is observed.

Optical registration, because of its high natural frequency and sensitiveness, shows other oscillations which are of lesser amplitude but which occur at definite moments of the cycle, *e.g.*, vibrations due to the second heart sound.

The *a* wave coincides with, and is due to, auricular systole. It begins at the same time as auricular contraction, but the end of auricular systole is not always clearly marked in the phlebogram. The descending limb of *a* begins when the first muscular units of the auricle relax, and continues to fall until all the fibers are relaxed. The end of this descending limb coincides with the intersystolic period (when there is one) and passes without transition to the oscillations which occur during isometric contraction of the ventricles. These oscillations are not very typical nor prominent.

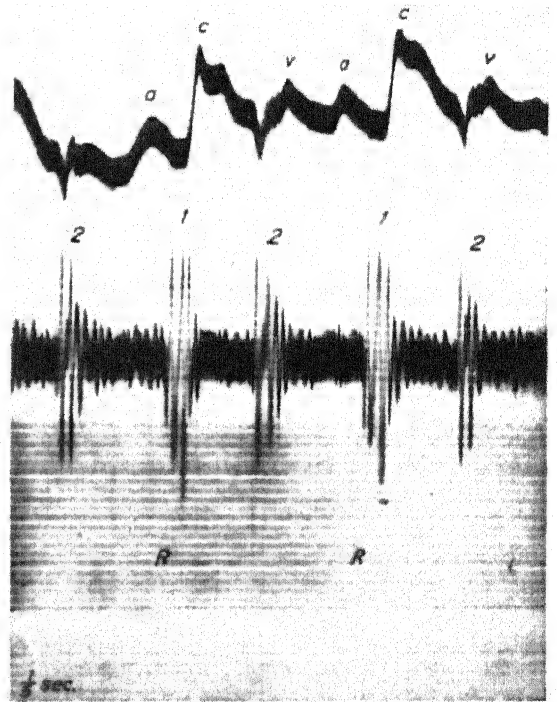



Fig. 120. "Arterialized" venous pulse. Phlebogram, phonocardiogram obtained by direct optical registration, and ECG; simultaneous records from a normal man. The *c* wave is prominent and similar to the central arterial pulse. (*Caeiro, 1942.*)

The *a* wave is absent in the phlebogram of cases of auricular fibrillation (Fig. 121). In auricular flutter the phlebogram consists of a series of *a* waves, interspersed periodically by the *c* and *v* waves of ventricular activity (Fig. 75).

The origin of the *c* wave has been the object of much discussion. Optical records undoubtedly

[1] Caeiro, A., *Medicina, Buenos Aires*, 2, 257, 1942.

show this wave to be due to an increase in venous pressure occurring during the ejection phase. At the same time intra-auricular pressure falls because of three concurrent circumstances: (*a*) the auricular muscle relaxes; (*b*) the A-V septum descends rapidly, because of ventricular contraction, which exerts a "sucking" effect on the auricle; (*c*) intrathoracic pressure falls, because the blood ejected by the ventricle sends blood out of the thorax, and this fall in pressure is transmitted to the auricles through its flaccid walls.

The intra-auricular pressure curve, therefore, shows no wave corresponding to the *c* wave of the phlebogram; consequently the latter must be of extracardiac origin. Mackenzie[1] thought this wave was due to an impact of the carotid artery and for this reason called it the *c* (carotid) wave. Wiggers[2] made careful comparative studies of the pressures in the right auricle, the veins, and the aorta, with the venous pulse, in simultaneous records with optical-registration methods of adequate efficiency. He concluded that the *c* wave is due to an arterial impact, but not necessarily to that of the carotid, because it persists after the carotid has been extirpated.

The carotid pulse, nevertheless, is partially registered in an ordinary phlebogram, and the *c* wave frequently shows the characteristics of the carotid pulse wave optically recorded. The *c* wave is less prominent as the record becomes a more exclusively venous pulse; if this wave is scarcely noticeable, the phlebogram can be called a pure venous pulse. In arterialized venous pulse records, the *c* wave shows some of the features of the central arterial pulse, and in mixed venous pulse records the appearance of *c* is intermediate between the two former cases (Figs. 119 and 120).

The depression *x*, following the *c* wave, is due to the fall in auricular pressure during the ejection phase of ventricular systole. It is more marked when the arterial impact is small, because the latter tends to suppress it. On the contrary, strongly arterialized phlebograms show almost no *x* depression.

The *v* wave usually begins at the end of the ejection phase.

If there is a very prominent *c* wave, the *v* wave begins in the diastole. The ascending limb of *v* is due to the progressive accumulation of blood in the auricle and veins while the A-V valves remain closed. The top (usually acute) marks, and is due to, the opening of the A-V valves. The veins are emptied by the flow of blood into the ventricles. The descending limb of *v* corresponds, therefore, to the rapid filling of the ventricle.

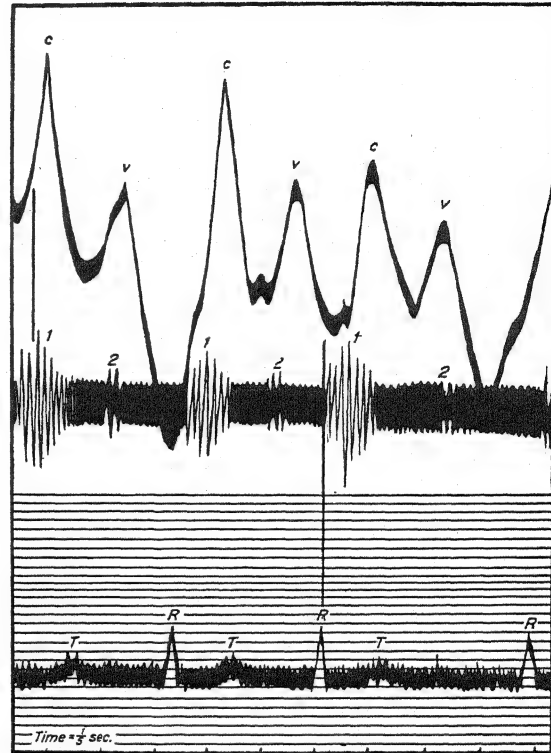



FIG. 121. Venous pulse in auricular fibrillation. Phlebogram, phonocardiogram, and ECG in a case of auricular fibrillation. The *a* wave of the phlebogram and the P wave in the ECG are missing. (*Orías, O., and E. Braun Menéndez, 1937.*)

On the ascending limb of the *v* wave in optical records there is always a notch, which corresponds to the second heart sound. If the segment capsule is very sensitive there will be two or three oscillations instead of a notch. This record of the second heart sound is very useful because it marks fairly precisely the end of the ejection phase and the beginning of isometric relaxation, which ends at the apex of the *v* wave.

The depression *y* is formed by the descending limb of *v* and the following rise in the record. This rise is due to the accumulation of blood in

[1] MACKENZIE, J., *J. Path. & Bact.*, 1, 53, 1893.
[2] WIGGERS, C. J., "The Pressure Pulses in the Cardiovascular System," Longmans, New York, 1928.

the auricle and veins after the ventricle has been filled and to the gradual increase in volume of the veins up to the beginning of the next cycle. If the diastolic pause is lengthened, a fourth wave, *h*, is observed. This wave occurs when there is no cardiac activity, and if the diastole is even more prolonged, other similar waves are seen. All these are of low frequency and of decreasing amplitude. At this moment of the cardiac cycle, ventricle, auricle, and veins form one large venous reservoir. The *h* and similar waves have been interpreted as vibrations of this reservoir caused by the reflection of the wave due to the rapid inflow into the ventricles on the opening of the A-V valves (Caeiro).

The *x* depression is frequently called the systolic collapse of the venous pulse, and the *y* depression the diastolic collapse.

Velocity of transmission of the venous pulse. The waves of the venous pulse are transmitted at much lower speed than the waves of the arterial pulse. Moreover there are great differences in the velocity of transmission of the different waves. These are due to differences in the mechanism of production, in the degree of filling of the veins, and in the velocity of the blood at the moment each wave appears.

In the dog, according to Caeiro,[1] the beginning of *v* spreads very slowly (0.85 m. per sec.) and the apex is transmitted at much higher velocity (3.86 m. per sec.). The *a* and *c* waves spread at velocities of 1.40 and 1.70 m. per sec. respectively. The oscillations of the second heart sound spread with a velocity of 4.70 m. per sec.

The heart cycle as portrayed in the phlebogram. The following phases of the cardiac cycle can be delimited fairly precisely in the phlebogram: auricular systole, or presystole (*a* wave); ventricular ejection phase (from the beginning of *c* to the oscillations of the second heart sound); isometric relaxation (from the second heart sound to the top of *v*); rapid filling of the ventricle (from the top of *v*, caused by opening the A-V valves, to the bottom of *y*); and diastasis (from the bottom of *y* to the beginning of *a* in the next cycle) (Fig. 114).

The A-V conduction time is marked in the phlebogram by the interval between the beginnings of *a* and *c*; normally it is not more than 0.20 sec. The *a-c* interval in the phlebogram is the equivalent of the PR interval in the electrocardiogram and, before Einthoven's work with the string galvanometer, was the only means of measuring A-V conduction time.

[1] Caeiro, A., *Rev. argent. de cardiol.*, **8**, 329, 1941.

The phlebogram is the record that permits the recognition of the greatest number of phases in the cardiac cycle. It affords the simplest and surest method of measuring the duration of the diastolic phases. Finally, as a reference tracing it is of great value for the precise localization in the cardiac cycle of acoustic and other signs of cardiac activity.

Some observers have tried to assess the dynamic conditions of cardiac activity and the distensibility of the heart walls by a minute analysis of the waves of the phlebogram, especially of their amplitude. The legitimacy of these conclusions is questionable, since in the same subject the records vary considerably according to the pressure exerted on the exploring funnel or its position with regard to the vein, without any reference to the conditions of cardiac activity.

BIBLIOGRAPHY

Burch, G. E., "A Primer of Venous Pressure," Lea & Febiger, Philadelphia, 1950.

Caeiro, A., "El Pulso Venoso Normal," Sebastián de Amorrortu, Buenos Aires, 1942.

Condorelli, L., Fisiopatologiá della Circolazione Venosa, V. and G. Pagano, Cattania, 1947.

Eyster, J. A. E., "The Clinical Aspects of Venous Pressure," Macmillan, New York, 1929.

Franklin, K. J., "A Monograph on Veins," Charles C Thomas, Springfield, Ill., 1937.

Golwitzer-Meier, K., *Ergebn. d. Physiol.*, **34**, 1145, 1932 (venous system).

Groedel, F. M., "The Venous Pulse and Its Graphic Recording," Brooklyn Medical Press, New York, 1946.

Houssay, B. A., "El Pulso Venoso," Buenos Aires, 1916.

Landis, E. M., and J. C. Ortenstine, Functional Significance of Venous Blood Pressure, *Physiol. Rev.*, **30**, 1, 1950.

Ochsner, A., R. Colp, and G. E. Burch, *Circulation*, **3**, 674, 1951 (venous pressure in normal persons).

Ohm, R., *Ztschr. f. exper. Path. u. Therap.*, **9**, 443, 1911 (venous pulse).

Orías, O., "Registro e Interpretación de la Actividad Cardíaca," 3d ed., El Ateneo, Buenos Aires, 1939.

Straub, H., *Deutsch. Arch. f. klin. Med.*, **133**, 253, 1920 (venous pulse).

Wiggers, C. J., "Pressure Pulses in the Cardiovascular System," Longmans, New York, 1928.

CHAPTER 24

Special Vascular Circuits

THE CIRCULATION THROUGH certain parts of the organism, because of morphologic and functional characteristics, deserves special consideration. One such circuit, the coronary system of the heart, has already been dealt with (see Chap. 18). Others will now be studied.

PULMONARY CIRCULATION

The pulmonary circuit consists of the pulmonary arteries, capillaries, and veins. The blood returned by the systemic veins to the right auricle is ejected into this circuit by the right ventricle. As it passes through the pulmonary capillaries the blood comes into close contact with the alveolar air, and the process of hematosis takes place, *i.e.*, CO_2 is lost and oxygen is taken up by the blood. The arterialized blood is then carried to the left auricle by the pulmonary veins. In the pulmonary circuit the arteries have thinner and less elastic walls than in the aortic circuit; they are also shorter and branch out more profusely. The peripheral resistance is much less than in the general circulation.

Pulmonary capillaries form a kind of hood closely fitted on the alveolar walls. Microscopic observation by transillumination shows the blood flowing through them continuously without pulsation, the erythrocytes passing in single file within the narrow capillaries.

Between the pulmonary and bronchial circuits (the latter branching off from the aortic system) there are only capillary anastomoses, without functional significance in normal conditions.

The pulmonary circuit has been explored directly in animals and man by introducing a catheter through the right ventricle[1] or by transthoracic puncture of the pulmonary artery.[1] Records of the blood pressure at different levels of the pulmonary artery can thus be obtained by connecting the catheter with an adequate manometer. In the main branches of the pulmonary artery the records show a similar contour to that of pressure records of the aorta, but at lower levels (Fig. 122). If the catheter is introduced until it can no longer advance (because it fills, and therefore occludes, the lumen of the vessel), a record is obtained, called the *capillary pulmonary pressure*,[2] which registers by retrograde transmission the pressure in the obstructed territory distal to the end of the catheter

In healthy subjects lying on their backs, the

FIG. 122. Human aortic and pulmonary pressures. Aortic pressure, pulmonary pressure, ECG, and time in 0.2 sec. (*After Cournand, A., Circulation, vol. 2, p. 641, 1950.*)

[1] COURNAND, A., H. D. LAWSON, R. A. BLOOMFIELD, E. S. BREED, and E. DE F. BALDWIN, *Proc. Soc. Exper. Biol. & Med.*, **55**, 34, 1944.

[1] VACAREZZA, R. F., A. LANARI, and A. J. ALBERTI, *Rev. argent. de cardiol.*, **13**, 205, 1946.

[2] HELLEMS, H. K., F. W. HAYNES, and L. DEXTER, *J. Applied Physiol.*, **2**, 24, 1949.

pulmonary circulation has the following characteristics:[1] minute volume, 5.5 liters/min., *i.e.*, 3.1 liters/min. per sq. m. of body surface; systolic pressure in the main branches of the pulmonary artery, 22 ± 2.9 mm. Hg; diastolic, 8 ± 1.7; mean, 13 ± 2.3; capillary pulmonary pressure, 5 mm. Hg.

The average oncotic pressure of plasma proteins is 25 mm. Hg; therefore water cannot filter through the capillary endothelium into the alveoli, and normally the alveolar cavity is free from fluid.[2] If pulmonary capillary pressure increases to a figure above the oncotic pressure of plasma proteins (*e.g.*, in left ventricular insufficiency) or if the capillary endothelium is damaged (*e.g.*, by toxic gases), fluid passes out of the capillaries and floods the alveoli, which become useless for hematosis (pulmonary edema).

The resistance of the pulmonary circuit is so low that in dogs[3] it is possible to destroy 90 per cent of the muscle of the right ventricle without causing changes in the pulmonary, aortic, or venous pressure. Apparently tension arising passively in the inactivated right ventricle owing to the contraction of the left ventricle is sufficient to convert the right ventricle into an efficient pump capable of maintaining a normal pulmonary circulation.

Recent work on the much-discussed question of the regulation of the diameter of pulmonary vessels by vasomotor nerves seems to prove that this type of regulation does not exist.[4]

The pulmonary circuit can act as a blood reservoir and take part in the regulation of the circulating blood volume.

Drinker and his collaborators[5] have demonstrated that storage of blood in the lung takes place at the expense of the air space. Changes in vital capacity observed, according to whether a subject is standing up, sitting, or lying down, are due to variations in the amount of blood stored in the lung in the different positions.

[1] Cournand, A., *Circulation*, **2**, 641, 1950.

[2] Wiggers, C. J., "Physiology in Health and Disease," 3d ed., Lea & Febiger, Philadelphia, 1939, p. 687.

[3] Starr, J., W. A. Jeffers, and R. H. Meade, Jr., *Am. Heart J.*, **26**, 291, 1943; Bakos, A. C. P., *Circulation*, **1**, 724, 1950; Kagan, A., *Circulation*, **5**, 816, 1952.

[4] Cournand, A., *Bull. New York Acad. Med.*, **23**, 27, 1947.

[5] Drinker, C. K., F. W. Peabody, and H. C. Blumgart, *J. Exper. Med.*, **35**, 77, 1922.

Redistribution of blood after partial pulmonary collapse, pathologic pulmonary processes, or surgical removal of part of the lung depends on local mechanical conditions. When there is a local increase in peripheral resistance limited to one lobe or part of a lobe, blood is shunted toward the intact parts of the lung. The large capacity of the pulmonary vascular bed facilitates redistribution. Thus in patients with only one lung, in which twice the normal amount of blood circulates, right ventricular pressure does not increase (Cournand).

CEREBRAL CIRCULATION

The nutrition of the brain is assured by the internal carotid and vertebral arteries, which anastomose to form the circle of Willis. This circle gives out arterial trunks which ramify and form the arterial network of the pia. From this network terminal arterioles penetrate into the depth of the brain. Occlusion of one of these arteries causes necrosis of the tissues it serves; "softening" thus produced disturbs the functions of the damaged centers and tracts.

The blood circulates in the brain through an incompressible mass enclosed in a cavity with inextensible walls (the cranium); therefore any amount of blood coming in by the arteries must cause the evacuation of an equivalent amount through the veins. This fact was pointed out by Monro in 1783 and is now known as the Monro-Kellie principle. Consequently, circulatory changes in the brain are brought about more by modification of the velocity of flow than by changes in the diameter of the blood vessels.

The cerebral blood vessels are nevertheless provided with vasomotor nerves. Stimulation of the cervical sympathetic produces all the signs of vasoconstriction in the encephalic mass of the corresponding side. Stimulation of the vagus, on the contrary, provokes vasodilatation.

Histamine dilates the vessels of the pia and increases their pulsation, causing headache. Local application of adrenaline provokes vasoconstriction, but injection of a dose sufficient to produce general effects causes the blood vessels of the brain to dilate passively, because the rise in blood pressure produced by vasoconstriction (especially in the splanchnic area and the skin) sends more blood into them. Ether and CO_2 act locally on the cerebral vessels and cause them to dilate.

The cerebral minute volume can be determined by the method proposed by Kety and Schmidt.[1] The subject breathes nitrous oxide so as to accumulate this gas in the blood at a concentration below the anesthetic level, and the Fick principle is applied in order to establish the amount of gas taken up by the tissues. The concentration of nitrous oxide in (*a*) arterial blood, (*b*) cerebral venous blood, and (*c*) cerebral tissue, is determined in samples drawn from the radial or femoral artery and the bulb of the internal jugular vein. The concentration in cerebral tissue is deduced from the concentration in venous blood when an equilibrium is established between the concentration of nitrous oxide in blood and in the tissues (it takes 10 min. to establish this equilibrium). With these data it is possible to calculate the cerebral minute volume. To determine the oxygen or glucose consumption of the brain, the arteriovenous difference in oxygen or glucose concentration is multiplied by the minute volume. Cerebral vascular resistance is equal to the ratio between the mean arterial pressure and the cerebral minute volume.

Normal subjects[2] have an average cerebral minute volume of 65 ml. of blood per min. per 100 gm. brain tissue, *i.e.*, approximately 14 per cent of the total cardiac output. The oxygen and glucose consumption are 3.8 ml. of oxygen and 6.2 mg. of glucose per min. per 100 gm. of brain tissue.

HEPATIC CIRCULATION. PORTAL SYSTEM

The circulation through the liver is peculiar in many respects. After birth the liver receives blood from two sources, the portal vein and the hepatic artery. The portal vein carries 75 per cent of the blood coming to the liver, but the 25 per cent brought by the hepatic artery is indispensable; if it is suppressed by tying this artery, the liver degenerates rapidly and the subject dies.

Treatment with large doses of penicillin prevents, in a high proportion of the cases, the hepatic necrosis thus provoked.[3] This finding suggests that necrosis is due to proliferation of anaerobic germs, which penicillin inhibits. The main role of the hepatic artery would therefore be to maintain a sufficiently high oxygen pressure to prevent growth of anaerobic bacteria, which are usually found in the phagocytic Kupffer cells after they have been absorbed from the intestine.

[1] Kety, S. S., and C. F. Schmidt, *Am. J. Physiol.*, **143**, 53, 1945; *J. Clin. Investigation*, **27**, 476, 1948.

[2] Scheinberg, P., and H. W. Jayne, *Circulation*, **5**, 225, 1952.

[3] Markowitz, J., A. Rappaport, and A. C. Scott, *Am. J. Digest. Dis.*, **16**, 344, 1949; Fraser, D., *et al.*, *Surgery*, **30**, 624, 1951.

These findings have led to the treatment of hepatic cirrhosis by ligation of the hepatic artery. The procedure has been used successrully in emergency treatment of hemorrhage due to rupture of varicose veins in the esophagus caused by hepatic cirrhosis.[1]

After entering the liver the portal vein branches out profusely. The final divisions run in the interlobular spaces and form around the lobules the periportal capillary network, of which the capillaries of the hepatic artery are also part. The sinusoids arise in the periportal capillary network and radiate into the lobules between the hepatic-cell columns toward the central vein which opens into a sublobular branch of the suprahepatic vein, tributary of the inferior vena cava. The muscle layer of the suprahepatic veins thickens considerably at the level of its opening into the inferior vena cava, thus forming a sphincter which regulates the flow of hepatic blood into the inferior vena cava. In the dog, however, hepatic congestion seen in shock (anaphylactic, histamine, etc.) is due to a diffuse spasm of the entire hepatic venous side of the liver, rather than to constriction of a more or less circumscribed sphincter.[2]

Microcinematographic studies *in vivo* of the amphibian liver have shown that the circulation in the intralobular sinusoids is fairly complicated. Blood flow and pressure are controlled by microscopic sphincters placed along the course of the sinusoids and in their opening into the central vein.[3]

The pressure in the portal vein has been measured in experimental animals; it is from 8 to 12 mm. Hg. In the hepatic artery it is 120 to 130 mm. Hg.

The hepatic minute volume in man can be calculated by the following method: Urea concentration in hepatic blood is determined by collecting a sample

[1] Chenoweth, A. I., *Ann. Surg.*, **135**, 756, 1952.

[2] There is suggestive but so far inconclusive evidence that a similar mechanism may be present in other animals (cat, white rat) and in man (Thomas, W. D., and H. W. Essex, *Am. J. Physiol.*, **158**, 304, 1949).

[3] Knisely, M. H., E. H. Bloch, and L. Warner, *Kgl. Danske Vidensab. Selskab., Biol. Skr.*, **4**, No. 7, 1948.

by means of a catheter introduced into an arm vein, then through the superior vena cava and right auricle into the inferior vena cava. Urea concentration in arterial blood and the amount of urea excreted by the kidney in given time are also determined. The minute volume of blood flow through the liver is then calculated by applying Fick's principle.[1] The elimination of bromsulfonphthalein injected has also been used for this purpose, applying a similar procedure.

In 73 normal men the hepatic minute volume was found to be 1,580 ± 32.5 cc./sq. m., and 1,340 ± 46.6 cc./sq. m. in 18 normal women. Patients suffering from hepatic cirrhosis had significantly smaller hepatic minute volumes.[2] When the subject stands up the hepatic minute volume decreases 40 per cent.[3]

The distribution of the portal system is such that all the blood from the abdominal viscera, especially from the intestinal tract, must come into close contact with the hepatic cells, which considerably modify this blood. Some substances brought by the portal blood are changed and stored in the liver cells; others are manufactured by the liver and confer to the blood some of its essential properties, *e.g.*, fibrinogen and prothrombin. The large and variable capacity of the portal system makes it an important blood reservoir, which plays a major part in the regulation of the circulating blood volume.

Sudden occlusion of the portal vein causes hemodynamic disturbances which rapidly end in death. Blood accumulates in its tributaries and the rest of the organism suffers from anemia. The animal, it has been said, bleeds itself into its own abdominal viscera.

Gradual occlusion of the portal vein is observed in cases of cirrhosis of the liver. In this disease proliferation of fibrous connective tissue strangles the blood vessels in the liver, and the portal blood finds its way to the vena cava through anastomoses that are normally minute and not visible. These anastomoses develop gradually into large vessels and form a conspicuous venous collateral circulation, draining the blood from the abdominal viscera into the vena cava.

The double blood supply of the liver (portal vein and hepatic artery) and the behavior of the heart in hypoxia, which differs according to whether the liver is excluded or not from the circulation, have led Rein[1] to postulate two types of hepatic function: (*a*) the ingestive function, in which the portal system plays the main part, the object of this function being to take up and process nutritive substances; and (*b*) the regulation of oxidation, in which the hepatic artery plays the main part. According to Rein, the liver contributes to the relief of general or local oxygen deficiency. The heart, the organ most studied in this respect, in conditions of hypoxia is assisted by a substance (*hypoxie-lienin*) released from the spleen by a nervous mechanism when there is an insufficient oxygen supply. This substance, on passing through the liver, acquires properties similar to those of strophanthin and regulates oxidative mechanisms in the myocardium. Activation of *hypoxie-lienin* in the liver requires the blood supplied by the hepatic artery.

Extirpation of the liver has been performed experimentally and has given important information on hepatic functions. Special measures must be taken to prevent death from circulatory disturbances caused by the ligature of the portal vein and the inferior vena cava; otherwise the animal dies before functional changes due to the absence of the liver become evident. Mann and Magath[2] were the first to obtain satisfactory results. The operation is performed in three stages, separated by intervals of several days:

1. Surgical anastomosis of the inferior vena cava into the portal vein (an inverted Eck fistula; see below), in order to increase the flow and pressure in the portal system so that a collateral circulation will develop, draining the system into the superior and inferior venae cavae above the level of the liver.
2. Ligature of the portal vein immediately before its entrance into the liver, an operation which is now well tolerated, because of the collateral circulation already developed.
3. Removal of the liver, which is considerably atrophied as a consequence of the second operation.

Hepatectomy can be performed in a single operation if the portal vein is anastomosed to the inferior vena cava and the segments of the vena cava above

[1] MYERS, J. D., *J. Clin. Investigation*, **26**, 1130, 1947.
[2] BRADLEY, S. E., F. J. INGELFINGER, and G. P. BRADLEY, *Circulation*, **5**, 419, 1952.
[3] CULBERTSON, J. W., *et al.*, *J. Clin. Investigation*, **26**, 1178, 1947.

[1] REIN, H., *Naturwissenschaften*, **36**, 233 and 260, 1949.
[2] MANN, F. C., and T. B. MAGATH, *Am. J. M. Sc.*, **161**, 37, 1921; *Am. J. Physiol.*, **55**, 285, 1921.

and below the hepatic veins are joined by a Pyrex glass tube.[1]

Eck's fistula is an anastomosis of the portal vein into the inferior vena cava, with the aim of excluding the liver from the portal circulation. The results are incomplete and transitory.

SPLENIC CIRCULATION

The spleen plays a part in the regulation of the circulation, especially in certain circumstances. The stroma and venous sinuses of the spleen give it the aspect of a sponge soaked in blood. The activity of the smooth-muscle fibers of the spleen causes its blood content to vary. The spleen is a blood reservoir which can remove from the circulation, or eject into it, appreciable quantities of blood, according to the state of relaxation or contraction of its muscles. Adrenaline and sympathetic stimulation contract the muscles of the spleen and therefore cause it to eject blood into the circulation. During exercise, asphyxia, cold, etc., the spleen volume becomes smaller.

Blood passes through the spleen in vessels with continuous although very thin walls. These capillaries allow the passage into the extravascular spaces of the red pulp not only of plasma but also of small particles and even erythrocytes.[2]

Blood stored in the spleen undergoes a process of concentration; the proportion of erythrocytes to plasma increases. This concentration is such that when the spleen contracts and ejects blood into the circulation, the erythrocyte count increases. Thus the spleen takes part not only in the control of the circulating blood volume, but also in the regulation of the number and concentration of erythrocytes.

[1] Markowitz, J., and S. Soskin, *Proc. Soc. Exper. Biol. & Med.*, **25**, 7, 1927; *J. Lab. & Clin. Med.*, **16**, 382, 1921.
[2] Peck, H. M., and N. L. Hoerr, *Anat. Rec.*, **109**, 447, 1951.

CIRCULATION TIME[1]

The time taken by the blood to traverse a part or the whole of a vascular circuit is called the circulation time. It can be measured by several methods, some of which can be applied in man. In general, a substance is injected into a peripheral vein and the time it takes to arrive at another point of the vascular system is measured.

Table 19. Circulation Time in Healthy Persons

	Substance used	*Circuit*	*Indicator*	*Circulation time, sec.*
Fischberg *et al.*	Saccharine	Arm–tongue	Sweet taste	9–16
Robb and Weiss	Sodium cyanide	Arm–respiratory center	Respiratory stimulation	9–21
Hitzig	Ether	Arm–lung	Smell	3.5–8
Spier *et al.*	$MgSO_4$ and Ca gluconate	Arm–throat	Heat	7–22
Stanojevic	Lobeline	Arm–coughing center	Cough	10.6 (av.)
Bernstein and Simkins	$MgSO_4$	Arm–throat	Heat	7–17.8

Many substances have been used: dyes, electrolytes, substances with a peculiar taste or odor, functional stimulants (vasodilators, stimulants for the carotid sinus or the respiratory center, etc.), radioactive isotopes, etc. In the histamine method[2] an adequate dose is injected into a vein in the arm and the time taken for the face and neck to flush is measured. At the same time the subject perceives a metallic taste, which can also be taken as the end point of the test. Histamine provokes headache, and this fact has limited its use. Other substances have been proposed: fluorescein (dye); sodium dehydrocholate, known as decholin, which gives a bitter taste; CO_2, which is inhaled and stimulates the respiratory center (this method measures the circulation time of the lung–respiratory center circuit); sodium cyanide (respiratory stimulant); lobeline (stimulates coughing); saccharine (sweet taste); sodium sulfate (sensation of heat in pharynx); calcium gluconate (sensation of heat in mouth and bitter taste); ether (sensation of smell).

[1] Fischberg, A. M., W. M. Hitzig, and F. H. King, *Proc. Soc. Exper. Biol. & Med.*, **30**, 651, 1932; Robb, G. P., and S. Weiss, *Am. Heart J.*, **8**, 650, 1933; Hitzig, W. M., *Am. Heart J.*, **10**, 1080, 1935; Spier, L. C., I. S. Wright, and L. Saylor, *Am. Heart J.*, **12**, 511, 1936; Stanojevic, L., *Ztschr. f. Kreislauff.*, **30**, 521, 1938; Bernstein, M., and J. Simkins, *Am. Heart J.*, **17**, 218, 1939; Wall, H. C., *Am. Heart J.*, **18**, 228, 1939; Walz, L., and G. Zimmermann, *Ztschr. f. Kreislauff.*, **40**, 2, 1951.
[2] Weiss, S., G. P. Robb, and H. L. Blumgart, *Am. Heart J.*, **4**, 1, 1929.

Circulation times observed with different methods are given in Table 19.

Considerable variations are found even in normal subjects, because the circulation time is conditioned by many factors: functional condition of the myocardium, minute volume, basal metabolism, blood distribution in the body, tone of vascular muscles, blood volume, composition and viscosity of the blood, hemoglobin concentration, number of erythrocytes, etc. The circulation time is considerably increased whenever there is any circulatory impediment. In cases of decompensated heart failure, the circulation time may be increased up to three times the normal.

Table 20 summarizes Vierordt's experiments in several species on the time taken for the blood to make a complete circuit from jugular vein to jugular vein. In these experiments the blood must pass through the aortic and pulmonary circuits. As the table shows, the number of heartbeats needed for a complete circuit is almost the same in all species. A longer circulation time is found in longer circuits (*e.g.*, femoral vein to femoral vein). The difference, nevertheless, is not considerable because the

Table 20. Time of Complete Circuit Related to Heart Rate

Animal	*Complete circuit time, sec.*	*Heart rate, beats per min.*	*Number of beats for complete circuit*
Rabbit	7.46	210	26.1
Sheep	14.14	110	26.0
Dog	16.7	96	26.7
Horse	31.5	55	28.8
Man*	23.1	72	27.7

Source: Tigerstedt, R., "Physiologie des Kreislaufes," vol. 4, Walter de Gruyter & Company, Berlin and Leipzig, 1923, p. 59.

* Figures calculated supposing man's size as intermediate between those of the dog and the horse.

greater part of the circulation time is taken up in crossing the capillary network; the blood circulates in arteries and veins at such a speed that differences in length are not of much importance.

BIBLIOGRAPHY

Dock, W., The Clinical Significance of Some Peculiarities of the Circulation in the Kidneys, Liver, Lung and Heart, *New England J. Med.*, **236**, 733, 1947.

Pulmonary circulation

Alberti, B. A. J., "La Presión en la Arteria Pulmonar," El Ateneo, Buenos Aires, 1948.

Cournand, A., Recent Observations on the Dynamics of the Pulmonary Circulation, *Bull. New York Acad. Med.*, **23**, 27, 1947.

Cournand, A., Some Aspects of the Pulmonary Circulation in Normal Man and in Chronic Cardiopulmonary Diseases, *Circulation*, **2**, 641, 1950.

Daly, J. de B., Reactions of the Pulmonary and Bronchial Blood Vessels, *Physiol. Rev.*, **13**, 149, 1933.

Straub, H., in Bethe's "Handbuch der normalen und pathologischen Physiologie," 7–1, 243, Springer, Berlin, 1926 (pulmonary circulation).

Wiggers, C. J., The Regulation of the Pulmonary Circulation, *Physiol. Rev.*, **1**, 239, 1921.

Cerebral circulation

Cobb, S., and W. G. Lennon, Cerebral Circulation. Intrinsic Control and Clinical Phenomena, *Federation Proc.*, **3**, 151, 1944.

Gregg, D. E., and R. E. Shippley, Experimental Approaches to the Study of the Cerebral Circulation, *Federation Proc.*, **3**, 144, 1944.

Kenk, R., and M. L. Nall, Physiology of the Circulation of the Brain. An Annotated Bibliography, 1938–1948, *Physiol. Rev.*, suppl. No. 1, 1952.

Scheinberg, P., and H. W. Jayne, Factors Influencing Cerebral Blood Flow and Metabolism, *Circulation*, **5**, 225, 1952.

Schmidt, C. F., "The Cerebral Circulation in Health and Disease," Charles C Thomas, Springfield, Ill., 1950.

Hepatic circulation

Andrews, W. H. H., B. G. Maegraith, and C. E. M. Wenyon, Studies on the Liver Circulation, II, The Microanatomy of the Hepatic Circulation, *Ann. Trop. Med. Parasit.* **43**, No. 2, 1949.

Bollman, J. L., Liver and Bile, *Ann. Rev. Physiol.*, **5**, 321, 1943.

Hoffbauer, F. W., J. L. Bollman, and J. L. Grindlay, Factors Influencing Pressure in the Portal Vein, *Gastroenterology*, **16**, 194, 1950.

Markowitz, J., and A. M. Rappaport, The Hepatic Artery, *Physiol. Rev.*, **31**, 188, 1951.

Snyder, C. D., Recent Advances in the Knowledge of the Liver, *Physiol. Rev.*, **22**, 54, 1942.

Splenic circulation

Barcroft, J., *Ergebn. d. Physiol.*, **25**, 118, 1926 (spleen and splenic circulation).

Peck, H. M., and N. L. Hoerr, The Intermediary Circulation in the Red Pulp of the Mouse Spleen, *Anat. Rec.*, **109**, 447, 1951.

CHAPTER 25

The Lymphatic System. The Cerebrospinal Fluid

THE LYMPHATIC SYSTEM

The lymphatic system is morphologically and functionally accessory to the blood vascular system. Its principal function is to carry back to the blood those products of cellular metabolism, or foreign substances, which cannot be reabsorbed into the blood capillaries. Some of the functions of the lymphatic system, such as the part it plays in immunity and in intestinal absorption, have taken on peculiar features, but these are only a special development of its general function of scavenging and draining the tissue spaces. The lymphocytes and monocytes of the blood are formed in the lymph nodes.

Gaspar Aselli, professor of anatomy in the University of Pavia, is credited with the discovery of the lymphatics. In 1622 he enthusiastically drew the attention of his contemporaries to these vessels, which are easily seen in the mesentery of a dog in the course of digestion, when they are filled with a milky fluid. Influenced by Galen's theories, he thought the lymphatics carried substances to the liver for the formation of the blood. A few years later (1651) Pecquet, a French physician, described the receptaculum chyli, also known as the cistern of Pecquet, and the thoracic duct with its opening into the veins in the neck. Later Rudbeck and Thomas Bartholin, who was the first to call these vessels the lymphatics, found them in other organs and tissues. By 1700 the existence of a vast system of vessels, different from the blood vessels, draining all parts of the body into the veins, was well known.

Anatomical distribution. The lymphatic capillaries have their origin in the intercellular spaces. They have endothelial walls, which are sometimes extremely thin and can be seen only with great difficulty, but which are never missing. There is general agreement that lymphatic vessels form, at their origin, a closed system; their content is not in direct contact with the cells, nor does it mix with the tissue fluids (Fig. 123).

At their site of origin lymphatic vessels have numerous anastomoses with each other and form a vast network containing the tissue cells in its meshes. From the capillary network collecting trunks arise; they are of larger diameter, have thicker walls, and are provided with valves that prevent the lymph from flowing back to the tissues. These trunks flow into others of still larger diameter, and in the higher mammals they end in two principal vessels: (*a*) the thoracic duct, which collects the lymph from all the parts below the diaphragm, including the abdominal viscera, and from the left half of the body above the diaphragm; (*b*) the right lymphatic duct or lymphatic vein, which collects the lymph from the right side of the body above the diaphragm.

The thoracic duct is the most important of all the lymphatic vessels. It joins the venous system at the angle formed by the junction of the left subclavian and jugular veins. The right lymphatic duct ends in the corresponding venous junction, on the right side.

The lymphatic trunks are of much smaller diameter than the arterial and venous trunks. Their walls are much thinner and are made up of endothelium surrounded by collagenous fibers and a few muscle fibers. The thickness of the walls and of the muscular coat increases with the diameter of the vessels.

In their course the lymphatics pass through structures, known as lymph nodes, made up of a connective-tissue stroma which sustains a large number of lymphocytes, reticular cells, and fixed macrophages; the whole is surrounded by a connective-tissue capsule. The afferent lymphatics of the nodes flow into sinuous

clefts called lymphatic sinuses, and these in turn empty into the efferent lymphatics. Lymph flowing through a node literally filters through it. Lymphocytes and monocytes are formed in the lymph nodes. These cells eventually pass into the blood together with the lymph.

Fig. 123. Lymphatics of the myocardium. A large collecting vessel and small capillary network. Note the irregularity of the vessel walls and the numerous anastomoses. (*After Bock, H., Anat. Anz., vol. 27, p. 40, 1905.*)

Methods of study. There are two principal methods for studying the lymphatic system: (*a*) a cannula is introduced into a lymphatic trunk and the lymph is collected, observing its rate of flow, amount, chemical composition, etc.; (*b*) the tissues are cut and a dye is placed in the wound, so that the coloring matter penetrates into the lymphatics and renders them visible. Drinker and his collaborators, and others before them, have obtained valuable information using the former of these methods; MacMaster and his collaborators have used the latter.

The formation of lymph. Lymph is formed by filtration of tissue fluids into the lymphatic capillaries. Ludwig was the first to postulate the existence of this mechanism of lymph formation; his ideas were given further experimental support by Starling, and they have found ample confirmation in more recent work. Lymphatic capillary endothelium plays a passive part in this process. It is not yet known how colloid and larger particles pass into the lymphatics. Foreign substances which have found their way into the tissues have a tendency to enter the lymphatics, unless they consist of very large particles.

In normal conditions lymph is formed from the excess of tissue fluid filtered and not reabsorbed by the blood capillaries (Fig. 117). Lymph formation is almost continuous in the intestinal tract, liver, and heart. In other tissues, lymph is formed in appreciable quantities only at certain times, as when filtration from the blood capillaries increases because the cells become active, or the tissues are massaged, or warmed, etc. The mechanism of filtration through the blood-capillary endothelium has been discussed in Chap. 22.

The amount of lymph formed varies considerably in different parts of the body at different times. Crandall *et al.*[1] were able to collect all the lymph flowing from the thoracic duct in a patient with a lymphatic fistula in the neck. The flow in basal conditions amounted to 1.38 cc per hr. per kg. of body weight. Maximum flow was observed after a meal; it was then 4 cc per min. per kg. Abdominal massage, or ingestion of water, oil, or protein, increased the flow by 3.19 to 5.28 per cent above the basal level.

Lymphagogues are substances that provoke an increase in lymph formation. Heidenhain made a detailed study of these substances and divided

[1] *Gastroenterology*, 1, 1040, 1943.

them into two classes. Lymphagogues of the first class are relatively complex and of many different kinds; they are found in tissue extracts, commercial peptone, etc. Lymphagogues of the second class are relatively simple crystalloids such as salts, urea, glucose, etc.

All the lymphagogues increase lymph production by modifying the process of filtration through the blood capillaries. Lymphagogues of the first class probably damage the capillary endothelium; they act on all the capillaries, but with particular intensity on the liver capillaries. Their action on the skin capillaries causes extensive cutaneous eruptions, with formation of wheals. Histamine, which is found in most of these extracts, produces the same effect.

Copious injections of aqueous solutions of crystalloids dilute the blood and diminish the oncotic pressure of the plasma proteins. This causes an increase in filtration of fluid from the capillaries to the tissues and in the formation of lymph.

Lymphagogues of the first class provoke the formation of concentrated lymph with a high protein content. Lymphagogues of the second class form dilute lymph with little protein.

Chemical composition and properties of lymph. The chemical composition of lymph varies considerably from one lymphatic territory to another. Even lymph from the thoracic duct, which comes from many and diverse territories, varies in composition according to the physiologic circumstances. With the exception of lymph coming from the intestine, which contains fats and other substances absorbed from the digestive tract, lymph is a watery solution of the same substances as are found in blood plasma. Crystalloid electrolytes in lymph are the same as those of blood plasma and are found in almost the same concentration. The protein concentration, on the other hand, is lower and variable. Lymph flowing from the thoracic duct has 3.2 per cent protein. Lymph from the limbs has 2.3 per cent protein, but this can fall to 0.5 per cent or less in the course of vigorous exercise. If the pressure in the hepatic veins is increased by partial occlusion of the inferior vena cava, the protein concentration of the lymph in the thoracic duct rises almost to the level of that in the blood plasma.[1]

The proteins in the lymph are the same as those in blood plasma, and it is thought that they escape from the blood through the blood-capillary endothelium. This escape is probably due to the frequent disturbance in capillary permeability caused by ischemia and other factors that act occasionally on the endothelium. These proteins pass from the tissue fluids into the lymphatics and are thus returned to the blood.

Fibrinogen also passes into the lymph. It is responsible for the coagulation of the lymph in the same way as for blood coagulation.

Thoracic-duct lymph has a milky appearance because of its high fat content, which increases during intestinal absorption, especially if the meal is rich in fats. Approximately 60 per cent of the total fat absorbed goes to the blood through the lymphatics. This is due probably to the formation of fat droplets in the process of absorption by the intestinal epithelium.

Lymphocytes are almost the only cells found in lymph. Occasionally monocytes are seen, and even less frequently granulocytes, usually eosinophils. In certain circumstances erythrocytes are found in the lymph. Lymphocytes are the specific lymphatic cells, formed in the lymph nodes and other lymphatic structures; other cells enter the lymph accidentally. The number of lymphocytes in the lymph varies considerably according to the territory and the circumstances in which the lymph is collected. In man between 2,000 and 20,000 per cu. mm. have been found.

The specific gravity of the lymph is 1.016 to 1.023. Osmotic pressure and alkaline reserve are slightly greater than in blood plasma; the viscosity is slightly less.

The lymph flow. Lymph flows from the tissues, where it is formed, to the opening of the collecting trunks in the veins of the neck. Several factors are responsible for its motion, which is similar to the blood flow in the veins. Valves direct and assist the progress of the lymph, but frequently valvular insufficiency permits its reflux.

Certain species (amphibians, reptiles, a few birds such as the swan, stork, goose, and ostrich) have lymph hearts, which drive the lymph forward. In amphibians there are two anterior and two posterior lymph hearts, one at the root of each limb. They have striated muscles and are innervated by fibers coming directly from the spinal cord. If these fibers are cut or the spinal cord is destroyed, the lymph hearts

[1] Starling, E. H., "The Fluids of the Body," Constable, London, 1909.

cease to beat. Under normal conditions the lymph from the limb and neighboring lymphatic sacs flows into the corresponding lymphatic heart and, because of the activity of the latter, is propelled into a short efferent lymphatic trunk which ends in the nearest large vein. Valves in this trunk direct the flow of lymph toward the vein. Destruction or paralysis of the lymph hearts causes death in a few days. Body weight increases considerably, and the blood becomes greatly concentrated because of the accumulation of fluid in the tissues.[1] The circulation of the lymph and its return to the blood are essential for life and the activity of the lymph hearts, in the animals under consideration, is indispensable for this circulation. Capillary filtration in these animals, in normal conditions, is considerably greater than in mammals. Obstruction of the lymph circulation causes "plasmatic hemorrhage" into the lymphatic sacs, a condition similar to mammalian shock.

In mammals there are no special organs that propel the lymph. Lymphatic trunks have plain-muscle fibers, and even in the small peripheral lymph vessels there are slow, rhythmical, active movements, peristaltic in some parts. Heller[2] counted about 10 pulsations per minute. The valves assure the progress toward the veins of lymph displaced by the contraction of the lymphatic musculature. Thus every segment between two valves acts as a miniature heart.

Muscles in the lymphatics respond to electrical stimulation and are under nervous control. The mechanism of the nervous regulation of lymphatic circulation in mammals is still unknown.

Apart from the activity of the muscles in the lymphatic walls, other forces contribute efficiently to the circulation of the lymph. "Negative" intrathoracic pressure, due to the elastic recoil of the lung, exerts a permanent aspiratory effect on the lymphatic system in the same way as on the venous system. This effect increases at each inspiration. The movements of the diaphragm act by a double mechanism on the lymphatic circulation: (*a*) the diaphragm is an inspiratory muscle and therefore its contraction increases the negative thoracic pressure; (*b*) on contracting, the diaphragm increases abdominal pressure, and by direct or indirect compression

[1] Braun Menéndez, E., and V. G. Foglia, *Arch. internat. de pharmacodyn. et de thérap.*, 64, 273, 1940.

[2] Heller, A., *Centralbl. f. d. med. Wissensch.*, p. 545, 1869.

of the lymphatics, especially of Pecquet's cistern, it propels the lymph toward the thorax.

Contraction of skeletal muscles has the same effect on lymphatic as on venous circulation, because of the flaccidity of the lymphatic walls and the existence of valves. The lymphatic vessels are squeezed between the muscles and emptied toward the veins. Fluctuations in pressure caused by the pulsation of neighboring arteries have the same effect. MacMaster[1] observed that perfusion of a rabbit's ear through its artery with a pulsating stream caused the lymph to flow at a speed 15 to 20 times that observed when the same ear was perfused at a constant pressure kept at the maximum level of the pulsating pressure.

Gravity helps or hinders the circulation of lymph according to whether the lymphatic territory is placed above or below its opening into the veins.

Velocity of flow and pressure in the lymphatics. The flow of lymph is neither continuous nor of uniform velocity; at times it has a pendulum movement.

If sodium sulfindigotate (indigo carmine) is injected into the leg of a dog, the dye is seen in the thoracic duct 10 min. later.

A warm temperature, activity, and hyperemia increase the velocity of flow; rest diminishes it.

Pressure within the lymphatics diminishes progressively from the tissues to their opening in the veins. Few accurate determinations of lymphatic pressure have been reported. In the ear of a mouse kept perfectly still, direct measurements made with micromethods have shown the pressure to be 0.7 ± 0.3 cm. H_2O.[2] In the thoracic duct, near its opening into the veins of the neck, the pressure is a few centimeters below atmospheric pressure. This difference in pressure causes the lymph to flow from the tissues to the veins.

In cases of heart failure with edema, MacMaster has observed that lymph remains almost stationary in the lymphatics of the skin. In cases of nephrosis with edema, on the other hand, lymph flows at a higher velocity than in normal conditions.

Lymphatic occlusion. Elephantiasis. Complete occlusion of the thoracic duct causes the accumulation of a milky fluid in the pleural and peritoneal

[1] MacMaster, P. D., *Harvey Lect.*, 37, 227, 1942.

[2] *Ibid.*

cavities (chylous ascitis). In the superficial territories there are no prominent changes, because they are drained through the numerous anastomoses of the lymphatic network. Complete obstruction of the thoracic duct may be observed in cases of mediastinal tumor.

In the limbs, lymphatic occlusion produces abnormal signs only when nearly all the lymphatic trunks have been obstructed. In such cases, considerable and progressive edema is produced. At first it is soft and depressible, but later it increases in consistency and becomes hard because of the proliferation of fibrous tissue. The diseased limb can take on gigantic proportions. When the condition has progressed to this stage it is called "elephantiasis," because the limbs look like those of an elephant. Infestation by parasitic worm (*Filaria*) is a frequent cause of this disease in some Oriental countries. It can be reproduced in experimental animals by obstructing the lymphatics with silica or other inert powders (Drinker).

Drinker suggests that pneumoconiosis is a kind of pulmonary elephantiasis, in which inhaled particles of dust penetrate the lymphatics of the lung and are carried to the regional lymph nodes. If there is an excessive accumulation of these particles, lymphatic obstruction is inevitable.

The lymphatic system and immunity. The lymphatic system plays an important part in the processes of immunity. The lymphatics in the skin and the mucosae of the digestive, respiratory, and genitourinary systems are directly exposed to the action of agents in the environment that attack the integrity of the organism. The slightest damage to these tissues opens the lymphatics to foreign bodies, of which some are inert particles but others are active microorganisms.

These particles, whether inert or active, are carried by the lymph to the lymph nodes, where they come into contact with the phagocytes (fixed macrophages) of the lymphatic sinuses. These cells capture the foreign particles and prevent them from entering the blood stream.

Inert particles (coal, silica, etc.) accumulated in the lymph nodes eventually cause fibrous proliferation in these nodes. This process is especially prominent in the lymph nodes of the lung, because even in ordinary conditions of life the inspired air carries dust particles. It is even more prominent in workers in mines, stone quarries, and other industries in which there is a great deal of dust in the air. Marked fibrosis in a large number of lymph nodes causes the disease known as pneumoconiosis (Greek *πνεύμων*, lung, and *κονία*, dust), referred to in a preceding paragraph.

Innocuous microorganisms are digested by the phagocytes. Virulent microorganisms cause inflammation in the site of entry or the corresponding lymph node, or both. If the microorganisms are highly virulent, the first line of defense in the lymphatics is overcome and they pass into the blood, provoking general infection (sepsis).

The concentration of immune antibodies in the lymph entering and leaving a lymph node has been examined. These observations, together with others, have led to the belief that lymph nodes produce antibodies. The results are not sufficiently clear to have convinced all workers in this field, and the subject is still under discussion.

Malignant tumors also spread along the lymphatic system. Tumor cells released from the parent tumor are carried by the lymph and proliferate in the regional nodes, or pass into the blood stream and are distributed all over the body.

CEREBROSPINAL FLUID

The central nervous system (brain and spinal cord) is bathed by a special fluid known as the cerebrospinal fluid, which also fills the internal cavities, *i.e.*, the ventricles and ependymal canal.

The cerebrospinal fluid occupies a completely closed space, without any direct communication with the blood vessels or the lymphatic system. This space comprises two compartments: (*a*) the subarachnoid space; (*b*) the ventricular cavities and the ependymal canal. The former is external and the latter internal with respect to the nervous structures.

The subarachnoid space is situated between the visceral layer of the arachnoid and the pia, which is closely attached to the nervous tissue and dips into all the junctions and fissures. Numerous fine trabeculae join both membranes and give a spongy aspect to the subarachnoid space (Fig. 124). The walls of this space and the trabeculae are lined by a continuous layer of flat endothelial cells, similar to those of the intima of the blood vessels. Blood vessels in the subarachnoid space are surrounded by the arachnoid, which sometimes forms a complete "cuff" around them.

The subarachnoid space varies in size in different parts. In some places it is a narrow cleft; in others it

forms irregular cavities, which are in free communication with each other and with all the rest of the subarachnoid space and are known as cisternae. One of the largest of these cavities is the cisterna magna, placed between the medulla and the cerebellum. Another important cavity is formed below the spinal cord. It ends, together with the pia, at the level of the first lumbar vertebra, while the dura and arachnoid extend down to the second sacral segment. Lumbar puncture is the introduction of a needle into this space, through the skin and soft parts between two vertebrae, in order to withdraw cerebrospinal fluid or to inject a substance into it.

Spinal anesthesia is performed by injecting a local anesthetic (*e.g.*, procaine) into the subarachnoid space by lumbar puncture. By this means, lengthy operations on the abdomen and lower limbs can be performed painlessly and without the patient losing consciousness.

The ependymoventricular space evolves from the primitive neural canal of the embryo. It is made up of the ependymal canal, the fourth ventricle, the sylvian aqueduct, the third ventricle, and the lateral ventricles. All these cavities communicate with each other and are lined by the ependymal cells of ectodermic origin.

The two main compartments occupied by the cerebrospinal fluid communicate with each other at the level of the fourth ventricle through the foramina of Magendie and Luschka.

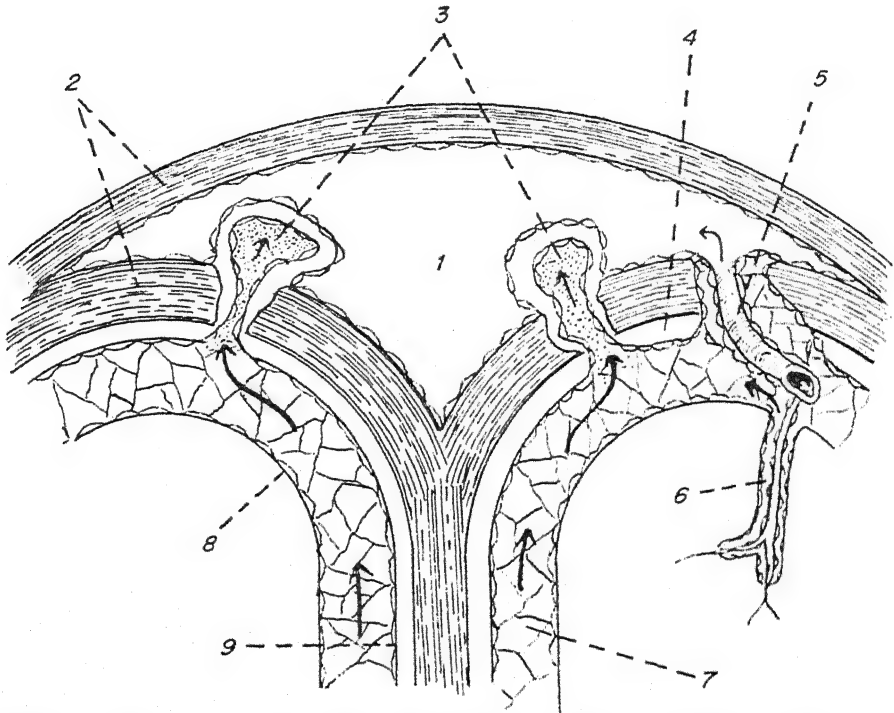



Fig. 124. Anatomical relations between the venous sinuses and the subarachnoid space. 1, venous sinus; 2, dura; 3, arachnoid villi; 4, subdural space; 5, cortical vein; 6, perivascular space; 7, subarachnoid space; 8, pia; 9, arachnoid. (*After Weed.*)

The blood vessels do not enter the subarachnoid space nor the internal cavities; the arachnoid or the ependymal membranes line their outer surface. Where an arteriole penetrates into the nervous tissue, the arachnoid and pia surround it and form a perivascular space, an extension of the subarachnoid space, which goes a short way into the depth of the nervous tissue (Fig. 124). In the ventricles the blood vessels in some places ramify profusely and form the chorioid plexuses, which seem to float in the ventricular cavity. They are covered by the ependymal lining and play an important part in the formation of the cerebrospinal fluid.

The subarachnoid space comes into intimate contact with the venous sinuses. In certain parts, fingerlike processes of the arachnoid, covered by the venous endothelium and the dura, project into the lumen of the sinus. The arachnoid villi (Fig. 124) thus formed

are submerged in venous blood, the subarachnoid space being separated from the blood stream by a very thin membrane. In the chorioid plexuses of the ventricles there is a similar situation, but in the opposite direction; the capillaries filled with blood are submerged in the cerebrospinal fluid, the blood being separated from the fluid by a very thin membrane. The chorioid plexuses play a fundamental part in the formation of cerebrospinal fluid, while the arachnoid villi do so in its reabsorption.

Table 21. Chemical Composition of Cerebrospinal Fluid

Inorganic substances	*Concentration, mg. per cent*	*Organic substances*	*Concentration, mg. per cent*
Sodium	335	Glucose	70
Potassium	106	Cholesterol	1
Calcium	53	Urea	15–25
Chloride	435	Proteins	10–25
Bicarbonate (CO_3H)	105	Total N	20
Phosphate (PO_4)	2	Fibrinogen	None

Chemical composition and physical properties. The total amount of cerebrospinal fluid has been calculated as 150 cc. This is the quantity of fluid that is found at a given moment, but it is continuously formed and reabsorbed, so that if a cannula is placed in the subarachnoid space much larger amounts can be collected. It is normally transparent and colorless; the specific gravity is 1.005. The protein content is only 18 mg. per cent, and as there is no fibrinogen, the fluid does not clot. Glucose is found in a concentration of 70 mg. per cent and sodium chloride about 700 mg. per cent. There are also small amounts of urea, creatine, cholesterol, etc. The average concentration of the principal substances in the cerebrospinal fluid is given in Table 21. There are few cells in the cerebrospinal fluid, mainly lymphocytes. The cell count is normally below 3 per cu. mm.

Formation and reabsorption of the cerebrospinal fluid. Cerebrospinal fluid is formed in the chorioid plexuses, the ventricles, and the perivascular spaces, especially those surrounding the capillaries of the pia. The process of formation is not well known; filtration and secretion seem to play a part in it.[1] Lymphagogues do not increase the formation of cerebrospinal fluid, but pilocarpine, which stimulates secretions, does; atropine and hyoscyamine suppress its formation.

[1] Obstruction of Monro's foramen provokes dilatation in the ventricle occluded; this does not occur if the chorioid plexus is extirpated.

There is a continuous flow of cerebrospinal fluid because of its incessant formation and reabsorption. Reabsorption takes place mainly through the arachnoid villi.[1]

The pressure of the cerebrospinal fluid is always higher than the pressure in the venous sinuses. Arterial blood pressure is an important factor in cerebrospinal fluid pressure, and reabsorption through the arachnoid villi is perhaps a simple process of filtration. Pressure in the cerebrospinal fluid, measured by lumbar puncture, is 6 to 12 cm. H_2O if the subject is lying down and 20 cm. H_2O if he is sitting up.

An obstruction in any part of the path followed by the fluid causes an increase of pressure upstream to the obstacle; therefore it continues to be formed even against pressures above those normally existing in the ventricles.[2]

Intravenous injection of hypertonic saline solution causes an immediate but transient rise in cerebrospinal fluid pressure, followed by a marked and prolonged fall. This is due to an increase in the osmotic pressure of the blood, which draws fluid from the tissues and cavities of the organism into the blood vessels. The escape of large quantities of fluid from the subarachnoid space causes the cerebrospinal fluid pressure to fall. The brain also loses water, and its mass diminishes. In neurosurgery hypertonic solutions are injected before opening the cranium so as to have more "space" for operating.

The cerebrospinal fluid is an important adjunct to the central nervous system, offering it support and adequate conditions for normal function.

BIBLIOGRAPHY

Athanasiu, J., and D. Calugareanu, in Richet, "Dictionnaire de physiologie," Librairie Félix Alcan, **10**, 475, Paris, 1928 (lymph).

Drinker, C. K., *Am. Heart J.*, **18**, 389, 1939.

Drinker, C. K., "The Lymphatic System," Stanford University Press, Stanford, California, 1942.

Drinker, C. K., and J. M. Yoffey, "Lymphatics, Lymph, and Lymphoid Tissue," Harvard University Press, Cambridge, Mass., 1941.

[1] Weed, L. H., *J. Med. Res.*, **26**, 21, 51, and 93, 1914.

[2] Obstruction of the sylvian aqueduct or of the foramina of Magendie and Luschka results in distention of all the ventricles (hydrocephalus).

FLEXNER, L. B., *Physiol. Rev.*, **14**, 161, 1934 (cerebrospinal fluid).

FREEMAN, S., and F. S. GRONDINS, *Surg., Gynec. & Obst.*, **73**, 195, 1941 (lymph).

GREENFIELD, J. J., and E. A. CARMICHAEL, "Cerebrospinal Fluid in Clinical Diagnosis," Macmillan, New York, 1935.

MACMASTER, P. D., *Harvey Lect.*, **37**, 227, 1942 (lymphatics).

MERRITT, H. H., and F. FREMONT-SMITH, "The Cerebrospinal Fluid," Saunders, Philadelphia, 1937.

ROUVIÈRE, H., "Anatomie des lymphatiques de l'homme," Masson et Cie, Paris, 1932.

SABIN, F. R., *Harvey Lect.*, **11**, 124, 1915–1916 (lymphatics).

SMITH, R. O., *J. Exper. Med.*, **90**, 497, 1949 (lymphatic contractility).

STARLING, E. H., "The Fluids of the Body," Constable, London, 1909.

WARREN, M. F., *Ann. Rev. Physiol.*, **2**, 109, 1940 (lymphatics).

WEED, L. H., *Physiol. Rev.*, **2**, 171, 1922; **13**, 80, 1933 (cerebrospinal fluid).

SECTION THREE

Respiration

Under conditions of aerobiosis, cells need oxygen for the conversion of the energy chemically bound in foodstuffs into work, heat, etc. Oxygen has the functions of a food and should be considered as such. Respiration is therefore a metabolic or nutritional function.

Oxidation takes place within the cells by a process known as internal respiration. In unicellular organisms, oxygen from the environment diffuses directly into the protoplasm, and the CO_2 formed by oxidation diffuses out of the cell. In multicellular organisms with differentiated tissues, respiration by simple diffusion is not possible; oxygen is carried by the blood to the tissue fluids which bathe the cells, but since it is not very soluble in water or plasma, special respiratory substances (*e.g.*, hemoglobin) have been developed that can combine loosely with the oxygen and so transport it in the blood to the tissues.

The blood is loaded with oxygen in special organs. These are the gills in aquatic animals like fishes; the digestive tract and skin in certain amphibians; the lungs in birds, mammals, and other animals that live in an atmosphere of air (external respiration).

The gaseous product of respiration, CO_2, is eliminated in higher animals by an even more complicated mechanism of transport than that which carries oxygen to the tissues.

Respiration in mammals comprises (*a*) pulmonary or external respiration, which includes the exchange of gases in the alveoli of the lungs, respiratory movements which bring about the renewal of air in the lungs, and the nervous and humoral mechanisms regulating these movements; (*b*) transport of oxygen and CO_2 by the blood, and the mechanisms that control it; (*c*) utilization of oxygen by the cells, or internal respiration.

The different aspects of this function are closely interdependent and are linked with other functions such as the circulation, the activity of the nervous system, and the acid-base equilibrium.

CHAPTER 26

The Mechanics of Respiration

GASES ARE EXCHANGED between the blood and air through the walls of the alveoli. The blood and air in the lungs must be continuously renewed in order to maintain efficient conditions for this interchange. Air is renewed by thoracic and pulmonary expansion, causing inspiration, and retraction, causing expiration. The anatomic and functional properties of the thorax are such that all three of its diameters are increased during inspiration.

Respiratory movements are caused by the active contraction of the respiratory muscles and the passive effect of the elasticity of bony structures and the lungs. Both these factors are essential for respiration. The movements of the respiratory muscles are controlled by the nervous system.

To obtain a clear idea of the mechanics of the thorax, it is best to consider separately the bony structures, the respiratory muscles, and the thoracic contents.

BONY STRUCTURES OF THE THORAX

The bony structures of the thorax are the thoracic vertebrae, the ribs with their cartilages, and the sternum. Each rib with its cartilage makes up a costal arch; each pair of costal arches, with its corresponding vertebrae and the sternum, forms a costal ring. The ribs increase in length from the first to the seventh, which is the longest; then decrease to the twelfth, which is the shortest. The anterior extremities of the last two ribs (the floating ribs) are unattached, and they do not take an active part in respiration. From a functional point of view, they form part of the abdominal wall.

The spinal column is extended and increases in length during inspiration, but to simplify the study of the mechanics of respiration, it can be considered as a rigid structure. It articulates with the ribs by two joints: one formed by the head of the rib, which articulates with two contiguous vertebrae, and the other formed by the articulation of the costal tuberosity with a transverse process. The distance between the centers of the two articular surfaces in the sixth rib is from 25 to 30 mm. Because of this anatomical arrangement the ribs have only one movement—that of rotation around an axis that unites the functional centers of both costovertebral joints (Fig. 125). Slight movements due to the elasticity of the ligaments or ribs can be discounted. The transverse and anteroposterior diameters of the chest are increased on contraction of the elevator muscles as a result of the direction of the axis of rotation, the form of the ribs, and their downward and forward position. Figures 125 and 126 show the changes in position of the sixth rib during inspiration and expiration. A better idea of these movements can be obtained by taking a rib (*e.g.*, the sixth), placing it in the position it would occupy in the thorax, then moving it slightly on its axis of rotation. It can easily be seen that the movement due to inspiration causes that part of the rib which is farthest from the mid-line to move even farther, while the anterior extremity moves forward.

The sternum moves with the ribs, and because of the greater motility of the middle ribs it is projected upward and forward during inspiration. The ribs do not all have the same degree of movement, because of their differences in length, form, and attachment to other anatomical structures.

The thorax can be compared to a truncated cone, which has a wider transverse than anteroposterior diameter. It can be divided functionally into three parts: the operculum, the

upper costal mechanism, and the lower costal or diaphragmatic mechanism.

The operculum is formed by the first pair of ribs and the manubrium. On inspiration the operculum rises, the anteroposterior diameter increases, and the manubrium is projected forward; the transverse diameter varies only slightly or not at all. Two important factors in this movement are the torsion of the first costal cartilage and the manubriosternal joint; this joint is movable until fifty or sixty years of age. The operculum acts especially on the apexes of the lungs.

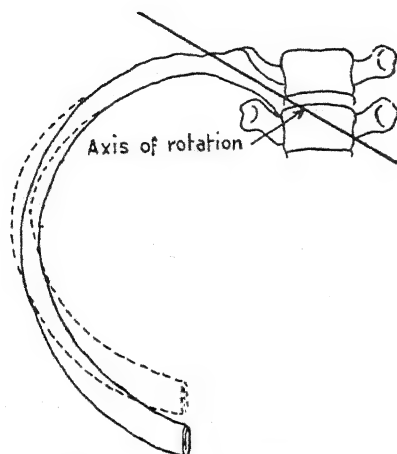



Fig. 125. Diagram showing movements of the sixth rib. The broken line indicates the position of the rib in inspiration; the solid line, its position in expiration.

The upper costal mechanism is formed by the costal rings corresponding to the second to fifth ribs and expands the upper part of the lungs. The lower costal mechanism is formed by the sixth to tenth ribs and expands mainly the inferior lobes of the lungs. As has already been mentioned, the last two ribs are not functionally part of the thorax.

RESPIRATORY MUSCLES

The contraction of the respiratory muscles causes active movement of the ribs, while passive movement is due to thoracic and pulmonary elasticity. Muscular contraction during respiration can be observed by simple inspection, by registration of the muscles' shortening by a myograph, or by a record of the electrical phenomena of their activity. The last, more modern, method is the most suitable, as it registers phenomena without traumatism, only requiring the introduction of needle electrodes into the muscles. This method demonstrates the progressive increase of nerve impulses during inspiration or expiration until they are suddenly interrupted by the initiation of the opposite act.

Inspiratory muscles. These muscles raise the ribs. They can be divided into principal muscles and accessory or emergency muscles. Only the former are used for normal inspiration; in forced inspiration the latter also take part.

The principal inspiratory muscles are the levatores costarum, the serrati postici superiores, the external intercostals, and the diaphragm. The first two groups are inserted on the spinal column and the ribs, and on contracting they raise the ribs. The external intercostals are inserted on two contiguous ribs and the corresponding cartilages; the fibers have a forward and downward direction. Their action is represented diagrammatically in Fig. 127. Records of the action potentials show that in quiet breathing the ribs are raised mainly by the contraction of the external intercostals and the interchondral part of the internal intercostals. Normal inspiration in the dog can be maintained by the contraction of these muscles alone.

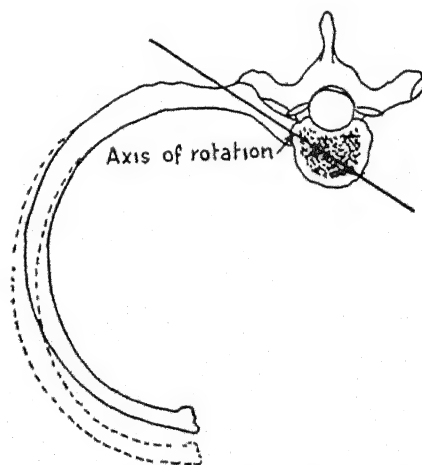



Fig. 126. The same diagram as in Fig. 125, seen along the vertical axis of the thorax.

The accessory inspiratory muscles act only in forced or dyspneic inspiration. This can be demonstrated by a simple observation: A thumb and finger are placed over the sternocleidomastoid muscles. During quiet breathing no change in tension is felt, but on making a sudden deep inspiration the contraction of the muscles

can be felt below the skin. These muscles and the scaleni raise the upper ribs.

The major and minor pectoral and serrati antici majores muscles raise the ribs when the shoulder and arm have been previously fixed. The shoulder girdle is fixed by the levatores angulorum scapulorum, trapezii, and rhomboidei. Patients in an attack of asthma or other dyspneic condition instinctively grip some firm object such as the head of the bed, so as to fix the upper limb and thus permit the action of the accessory respiratory muscles.

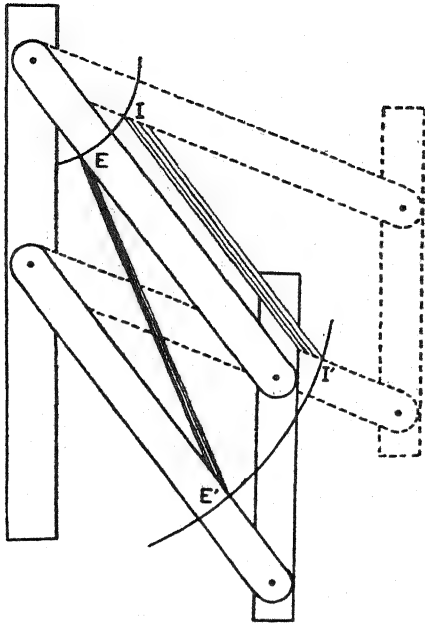



FIG. 127. Diagram showing the action of the external intercostal muscles. The vertical bars represent the vertebral column and the sternum; the oblique bars show the ribs in expiration (solid line) and in inspiration (broken line). The external intercostal muscles shorten on contraction and pass from the position *EE'* to *II'*, raising the ribs.

The diaphragm is the principal inspiratory muscle, and it can maintain adequate respiration when all the other respiratory muscles have been paralyzed. This can be shown in animals by cutting the spinal cord immediately below the origin of the phrenic nerve, an operation which paralyzes nearly all the other inspiratory muscles. Cases of section of the spinal cord at this level have demonstrated that in man also the diaphragm has a similarly predominant part in inspiration.

The central part of the diaphragm is a tendinous membrane. Its projection on the chest wall corresponds to the sixth chondrosternal joint. It is fixed to the mediastinum by the fibrous pericardium. Muscle fibers are inserted on all the periphery of this tendinous membrane. Some of these fibers have a backward and downward direction, and their other end is inserted on the lumbar vertebrae; they form the crura of the diaphragm. The rest of the muscle fibers form a dome-shaped structure descending to insertions on the lower ribs and the sternum. On the right side the diaphragm covers the upper surface of the liver; its highest point is projected on the chest wall at the level of the fourth intercostal space in the axillary line, in normal expiration. The left side is 14 to 27 mm. lower than the right (Figs. 128 and 129).

The dome shape is due mainly to the negative intrathoracic pressure which raises the diaphragm and draws it laterally against the thorax, closing the costophrenic angles. Abdominal pressure is an accessory factor. The diaphragm still keeps its domelike position in the cadaver after the abdomen has been opened and the viscera removed, but it becomes flaccid and is no longer dome-shaped if the thorax is also opened, causing the lungs to collapse. Muscular tone and active contraction of the diaphragm lower the dome. The mechanics of the diaphragm are not simple. The only fixed insertion of its fibers is that on the lumbar vertebrae. Contraction of the crura pulls the diaphragm down, lengthens the vertical diameter of the thorax, and causes the base and hilum of the lungs to descend. The other insertions are all made on mobile structures and it will be necessary to consider the results when these are fixed or mobile. When the tendinous center is fixed, the costosternal fibers raise the lower ribs, and like the other muscles that have the same effect on the upper ribs, they increase the transverse diameter of the thorax. If, on the contrary, the costosternal insertions are fixed, the diaphragmatic dome descends, and the vertical diameter is increased. In quiet breathing the diaphragm descends only slightly. Thus contraction of the diaphragm increases the three thoracic diameters.

The cross-sectional area of the thorax at the level of the diaphragmatic dome is approximately 250 sq. cm. If the movement of the diaphragm is compared to a piston, a descent

of 10 mm. will increase the thoracic volume by 250 cu. cm; *i.e.*, an amount equivalent to half the tidal volume (see "The amount of air in the lungs," page 250). This calculation is not accurate, because the movements of the diaphragm are not the same as those of a piston; nevertheless it gives an approximate idea of the importance of the diaphragm in respiration.

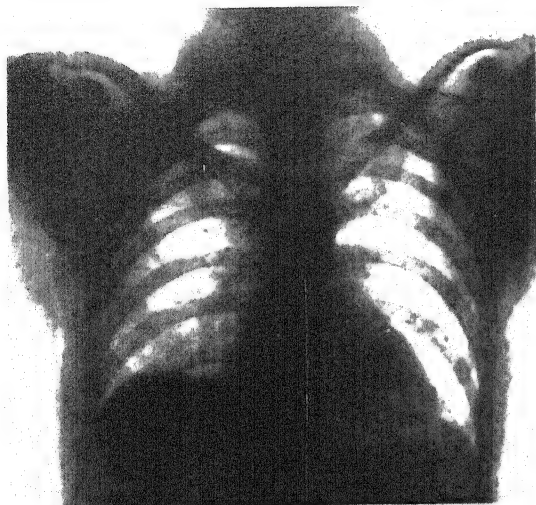



FIG. 128. X-ray of thorax in normal expiration.

As the diaphragm descends it exerts pressure on the abdominal viscera, which in their turn move the abdominal walls. Thus normal inspiration increases the diameters not only of the thorax but also those of the abdomen, and the type of respiration is known as "thoracic" or "abdominal" according to whether the contraction of the thoracic muscles or the diaphragm is most prominent.

In paralysis of the diaphragm it behaves as a passive membrane and ascends during inspiration; thus while the thorax widens, the abdominal girth diminishes. This is known as "paradoxical" respiration.

Expiratory muscles. The ribs are lowered by the contraction of the interosseous part of the internal intercostals serrati postici minores and the triangular muscle of the sternum. Contraction of the muscles of the abdominal wall compresses the viscera and by this means exerts pressure and raises the diaphragm. In normal quiet breathing expiration is mainly passive; thoracic and pulmonary elasticity brings the thorax and its contents back to a position of elastic equilibrium at the end of a normal expiration.

The costophrenic angles. Part of the diaphragm is pressed against the thoracic wall during expiration; the angle thus formed is the costophrenic angle. On contraction of the diaphragm the curvature of its dome tends to

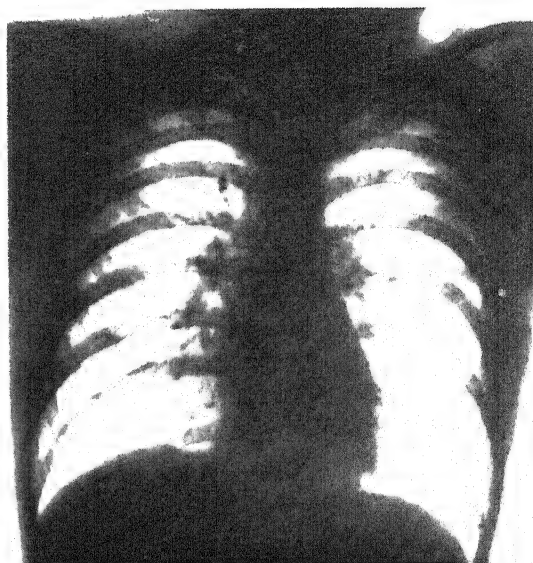



FIG. 129. X-ray of thorax of the same person in forced inspiration. Observe the increase in the transverse and vertical diameters of the thorax, the descent of the diaphragm, and the increase in the costophrenic angle.

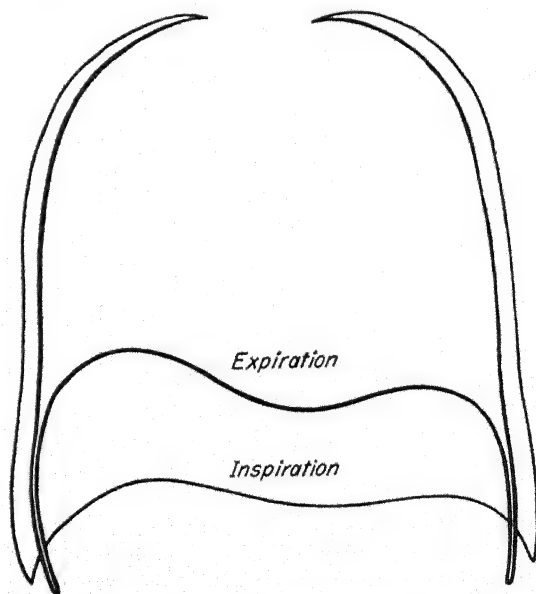



FIG. 130. Diagram of the anteroposterior projection of the costal and diaphragmatic pleura in expiration and in forced inspiration. Tracings made from the x-rays of Figs. 128 and 129.

diminish because of the shortening of the muscle fibers, thus separating the diaphragm from the thoracic wall. In expiration the diaphragm is pressed against the thoracic wall over an area of 6 to 8 cm. in width; in inspiration, as the diaphragm separates from the wall and the angle opens, the lung descends into it.

Radiography shows these changes in the position of the lung (Figs. 128, 129), which can also be demonstrated by percussion of the base of the thorax. The inferior limit of pulmonary resonance descends in forced inspiration by four fingerbreadths.

THE LUNGS

The lungs and other thoracic viscera (heart, blood vessels, etc.) are subject to passive changes caused by the expansion and retraction of the thoracic wall.

Thoracic and pulmonary elasticity. Elasticity is the capacity of a strained body to recover its shape and size after the external forces that strain it cease to act. Elasticity varies in different bodies, and if the strain exceeds it, the original shape is not recovered.

The thoracic wall is elastic; its resting position is that of normal expiration. The elastic force must be overcome during inspiration, so that expiration is mainly due to elastic rebound. In forced expiration the expiratory muscles must overcome the elastic force; in this case, the beginning of inspiration is accompanied by elastic rebound.

The lungs are elastic structures permanently distended under normal conditions, but if the thorax is opened and air penetrates into the pleural cavity, the lung collapses and is reduced to a small mass at the hilum. This is observed not only in living subjects but also after death.

Pulmonary elasticity can be measured in the human cadaver by placing a tightly fitting cannula in the trachea and connecting it with a manometer. When the pleural cavity is opened, the lung retracts and the pressure in the trachea rises 6 to 7 mm. Hg (*i.e.*, 80 to 95 mm. H_2O). If the initial position is that of expiration, the pressure rises only 3.9 mm. Hg (53 mm. H_2O); it rises 9.4 mm. Hg (130 mm. H_2O) if the lung has been distended in inspiration. The difference in elastic strain between inspiration and expira-

FIG. 131. Diagram demonstrating pulmonary elasticity. The elastic recoil of the lung tends to reduce the size of the lung in the sense indicated by the arrows. This produces a "negative" pressure in the pleural cavity, which is measured by puncturing the pleural cavity and connecting the needle with a manometer.

tion is therefore equivalent to a pressure of 5.5 mm. Hg. Pulmonary elasticity tends to reduce the volume of the lungs, and the inspiratory muscles must overcome it to distend them. Expiration, on the contrary, is favored by the pulmonary elastic force and in normal conditions is mainly passive.

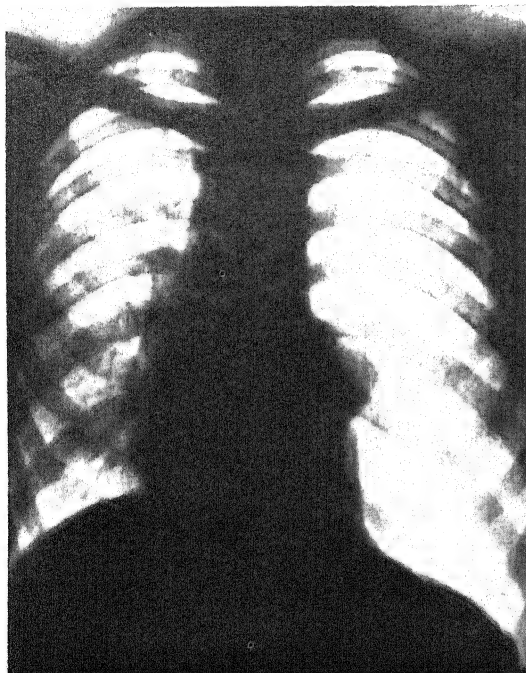



FIG. 132. X-ray of a case of pneumothorax. Air has penetrated into the left pleural cavity, the lung on that side has collapsed and is reduced to a small mass around the hilum. The heart is displaced toward the right. The diaphragm is lower on the left side than on the right because there is no "negative" pressure, owing to the absence of the elastic recoil of the lung.

Negative intrathoracic pressure. Pulmonary elasticity can also be demonstrated by puncturing the pleural cavity with a needle of wide bore and connecting it with a manometer (Fig. 131). Pressure in the pleural cavity is seen to be below atmospheric pressure and is usually called the "negative intrathoracic" or "negative intrapleural" pressure. This negative pressure is due to (*a*) the elasticity of the lungs; (*b*) the resistance with which the thoracic wall opposes this elasticity. Negative pressure is from 5 to 8 mm. Hg (70 to 110 mm. H_2O) in expiration and from 10 to 15 mm. Hg (135 to 200 mm. H_2O) in inspiration. Even in maximum expiration the lung is still submitted to elastic stress. Negative pressure maintains the domelike position of the diaphragm and also favors the entrance of blood into the large veins of the thorax.

Intrapleural pressure becomes positive (*i.e.*, above atmospheric pressure) when an expiratory effort is made while keeping the glottis closed, as in the act of defecation. This positive pressure causes several changes in the dynamics of the circulation.

Pneumothorax. If air penetrates into the pleural cavity accidentally or is introduced for the purpose of experiment or treatment, the lung collapses. The condition is known as pneumothorax (Fig. 132); it is called "open pneumothorax" when the wound in the lung or the thoracic wall remains open and lets the air circulate through it; otherwise it is called "closed pneumothorax." This condition causes severe disturbances in respiratory mechanics and in circulatory dynamics. The diaphragm is no longer drawn into the thorax by the elastic force of the lung, so it loses its domelike shape on the side corresponding to the pneumothorax (Fig. 132). The mediastinum is not a rigid structure, so it is pulled, together with the organs it contains, toward the normal side by the elastic effect of the normal lung (Fig. 132). The entrance of air is diminished not only in the collapsed lung, but also in the normal one, because its volume is decreased by the deviation of the mediastinum.

Unilateral pneumothorax is well tolerated in normal, quiet, physiologic conditions, and it is frequently provoked for the treatment of pulmonary tuberculosis. When there is anoxia such as that due to high altitude, decompensated cardiac disease, pneumonia, etc., pneumothorax has serious effects (dyspnea, cyanosis) and may even cause death.

In an open pneumothorax the severity of the condition depends on the size of the pleural opening. The ratio of air entering the lung to air entering the pleural cavity is the same as the ratio of the cross section of the trachea to the cross section of the pleural opening. If the cross section of the wound is equal to the cross section of the trachea, equal volumes of air enter the lung and the pleural cavity. In this case expansion of the thorax to twice that of normal inspiration will maintain a normal volume of tidal air. The size of the wound compatible with life will therefore depend on the vital capacity of the patient. If maximum expiration and inspiration

do not result in a normal volume of tidal air, the patient will be in danger of asphyxia. In subjects with a small vital capacity, an open pneumothorax is a more severe condition than in subjects with a large vital capacity. In the course of thoracic operations, artificial respiration by rhythmic insufflation should be kept up as long as the pleura is open; if only a small wound is made, however, this may not be necessary. Adhesions that bind the lung to the thoracic wall prevent the total collapse of the lung and diminish the severity of the symptoms of pneumothorax.

The decrease in the negative intrathoracic pressure hinders the entrance of blood into the large thoracic veins; therefore there is stasis in the extrathoracic veins, and the venous return to the heart diminishes.

An open pneumothorax causes considerable loss of heat and may be followed by a rapid fall in body temperature.

The movements of the lung. The total cross section of the respiratory tract increases progressively as the bronchi divide, so that it can be represented by a cone, with the apex in the trachea and the base in the alveoli. The total alveolar surface is approximately 100 sq. m.

The bronchi divide successively down to the terminal bronchioles, which have a smooth-muscle layer, the contraction and relaxation of which modify the amount of air going into or out of the alveoli. Terminal bronchioles branch out into respiratory bronchioles, so called because they have alveoli in their walls. The alveolar ducts open into the respiratory bronchioles. These ducts lead to the atria into which the air sacs open. The walls of the air sacs are made up of alveoli, but alveoli are also found in the walls of all the air tubes smaller than the terminal bronchioles (Fig. 133).

The expansion of the lungs is due to lengthening and dilatation of the fine branches of the air passages below the terminal bronchioles. The most distensible parts are the alveolar ducts; probably the alveoli do not change much in volume.

The lung is not dilated uniformly in all its parts during inspiration. The hilum, formed mostly by bronchi and blood vessels, has only slight distensibility. The middle of the lungs, where the bronchi and blood vessels ramify and where there is some parenchymatous tissue, is moderately distensible. Distensibility is at a maximum near the surface of the lung down to 2.5 to 3 cm. below this surface. As a whole the lung behaves more like a bellows than like a distensible sac.

The different parts of the lung do not expand to an equal extent in inspiration. The lower lobes expand more than the upper ones. The apex has not a very active part in respiration; in forced inspiration it is contracted and in forced expiration it is dilated. Two other regions, the mediastinal and dorsal surfaces, are not expanded much in inspiration. The diaphragmatic and costosternal surfaces expand most.

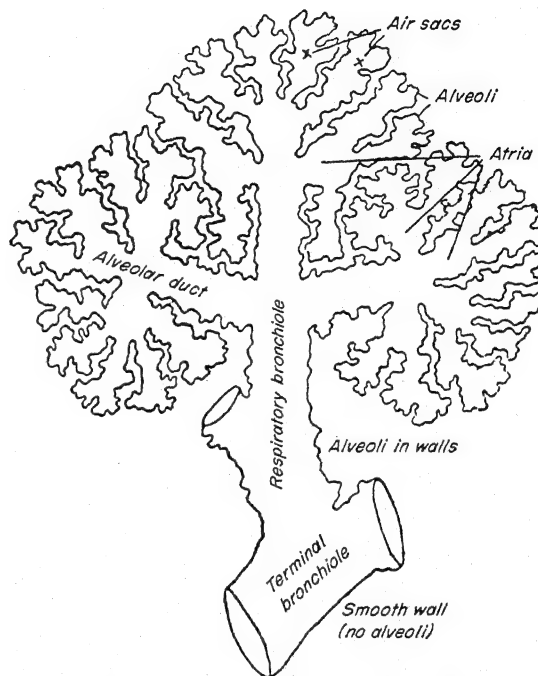



FIG. 133. Diagram of a pulmonary lobule. Three acini connected to a respiratory bronchiole are shown. The air sacs with the alveoli open into the atria and alveolar ducts. (*According to Miller.*)

The thoracic wall and the lungs move simultaneously, but the pulmonary and parietal pleurae slide on each other. The hilums move downward in inspiration, because of the descent of the diaphragm, and pull down the trachea. This can be easily seen by observing the descent of the larynx in inspiration. The displacement of the pulmonary surface varies in the different parts; it is not marked in the apexes, nor in the dorsal surface, but is considerable in the lower part of the lung, especially in the region of the

costophrenic angles, into which the lungs expand.

Pressure in the airways. The air passes in and out of the lung as a result of rhythmic changes in pressure, which can be easily measured and registered. In quiet breathing the difference in pressure between the atmosphere and the air in the lungs is only 2 to 4 mm. Hg, but in forced breathing it is much higher. A forced inspiration through a tube connected to a manometer can produce a negative pressure of −75 to −147 mm. Hg; a forced expiration can raise the pressure to +105 and even +256 mm. Hg. (These extreme figures are exceptional.) The maximum expiratory effort is greater than the maximum inspiratory effort. Expiration is used for voice production and to remove foreign particles from the respiratory tract by coughing, etc.

FIG. 134. Subdivisions of lung volume. (*Federation Proc., vol. 9, p. 602, 1950.*)

Circulatory changes caused by respiratory movements. Pressure in the pleural cavity is always negative, even in expiration, except when an expiratory effort is made keeping the glottis closed. This negative pressure assists the inflow of blood into the large thoracic veins and causes a negative pressure in the large veins of the neck. (There is danger of air embolism if these veins are wounded.) Negative pressure is greater in inspiration; therefore in inspiration more blood as well as more air flows into the lungs. This is an interesting example of functional coordination.

The amount of air in the lungs. (Fig. 134.) At a meeting held in 1950[1] the definitions of lung volumes and the terminology used were discussed in order to provide uniformity in the literature. It was recommended that the word "air" be replaced by "volume," "complemental" or "complementary air" by "inspiratory reserve volume," and "supplemental" or "reserve air" by "expiratory reserve volume," as these terms have a functional significance and are more easily remembered. A systematic set of symbols for the different variables in respiration physiology was also recommended. Subdivisions of the lung volume are represented in Fig. 134.

The following are the average figures for the lung volumes:

	Liters
Vital capacity:	
In men	3–4
In women	2–3
Inspiratory reserve volume, resting	2
Expiratory reserve volume, resting	1.5
Residual volume	1.5
Tidal volume	0.5

These volumes vary with the conditions of the subject, *e.g.*, rest, activity, etc. The residual volume varies with inspiratory or expiratory effort, partly because of changes in the amount of blood in the thorax, which increases with inspiratory effort and decreases with expiratory effort. The residual volume varies inversely to the amount of blood in the thorax.

The tidal volume is the air entering and leaving the lungs in a respiratory cycle. Of the total 500 cc., only about 360 cc. reaches the alveoli and takes part in hematosis. The rest (approximately 140 cc.) remains in what is called the dead space, *i.e.*, the large bronchi, trachea, larynx, and nasal cavities. When masks are used, *e.g.*, for anesthesia or protection against toxic gases, the volume of the dead space increases. The tidal volume must then be increased in order that sufficient fresh air may reach the alveoli, and therefore the pulmonary ventilation must be increased.

Changes in the volume of the dead space. The dead space is that part of the respiratory tract which has no alveoli. It therefore extends down to, and includes, the terminal bronchioles. The volume of air in the dead space varies, because the airways are not rigid structures. During inspiration the respiratory tubes are lengthened and dilated, especially in deep breathing.

The "physiologic" or "effective" dead space is the total space of the respiratory tract which, immediately before expiration, is filled with atmospheric air unmixed with alveolar air, *i.e.*, air that has not reached the region of

[1] *Federation Proc.*, **9**, 602, 1950.

gaseous exchange. The physiologic dead space can be calculated by subtracting the alveolar expired air from the total volume of expired air. Alveolar expired air is calculated from the concentration of CO_2 in expired and in alveolar air. For example, the total volume of expired air is 485 cc.; CO_2 concentration in expired air is 4 per cent and in alveolar air it is 6 per cent. The alveolar expired air in this case is

$$485 \times 4/6 = 323 \text{ cc.},$$

and the physiologic dead space is

$$485 - 323 = 162 \text{ cc.}$$

Spirometry. The volume of expired air and vital capacity are measured by instruments called spirometers, of which there are several models. Hutchinson's spirometer consists of a metal cylinder of adequate capacity, which is filled with water; a metal tube pierces the bottom of the cylinder, to which it is soldered, and rises above the level of water. This tube is connected by rubber tubing with a mouthpiece into which the subject breathes. The spirometer bell is another metal cylinder of slightly smaller diameter than the first one, into which it fits. It is completely submerged in the water (the air within it escapes through the metal tube) and its weight is accurately balanced by a counterweight. The subject breathes out by the mouthpiece into the spirometer; and the air raises the bell. The volume of air is read on a scale along which an indicator attached to the bell moves. To measure the air breathed out during several minutes, a system of valves is added so that atmospheric air is inspired and the expired air is collected in the spirometer or passed through a gas meter.

Vital capacity. The volume of air expired by a maximal expiration following a maximal inspiration has been given the name of vital capacity.[1] It is approximately 3 to 4 liters in men, exceptionally 5 to 6 liters; in women it is usually 2 to 3 liters.

The determination of the vital capacity is frequently used in armies and athletic clubs to estimate indirectly the capacity of an individual for physical effort. Vital capacity decreases when the elasticity of the thorax, or the lung, or both diminish (pleurisy, emphysema, sclerosis of the lung, ossification of the costal cartilages, etc.). Pain in the thorax or near the thorax (pleurisy, peritonitis) also diminishes vital capacity, because respiratory movements increase the intensity of pain. Several factors have an influence on vital capacity: *e.g.*, posture (it is greater when standing up than lying down), age (it increases up to twenty or thirty years, and later diminishes), sex, race, occupation, etc. It can be increased by training, and in athletes it is considerably above the average.

The relation between vital capacity (VC) and other bodily measurements has been established:

$$\frac{(\text{Weight})^{0.72}}{VC} = K_1$$

$$\frac{(\text{Sitting height})^2}{VC} = K_2$$

$$\frac{\text{Chest circumference}}{VC} = K_3$$

The average normal values for these constants are

$$K_1 = 0.69,\ K_2 = 1.85,\ K_3 = 2.41.$$

$$\frac{VC,\ \text{in liters}}{\text{Body surface, in sq. m.}} = 2.61\ (\text{men}),\ 2.07\ (\text{women})$$

Standing height multiplied by 22 to 24 gives the vital capacity in cubic centimeters in men; the height should be multiplied by 20 in women and 15 in children. Vital capacity is more closely related to height than to weight in fat and thin persons.

Altamira[1] has found the following average values in 431 healthy men eighteen to thirty-eight years old:

$$\begin{aligned}\text{Vital capacity} &= (\text{Body surface} \times 1{,}600) + 1{,}225\\ &= (\text{Standing height} \times 4{,}860) - 4{,}220\\ &= (\text{Weight} \times 18) + 2{,}900\\ &= (\text{Chest circumference} \times 42.5) + 440\end{aligned}$$

Of all these correlations the most reliable is that with the body surface.

Pulmonary ventilation. About one-third of the 500 cc. of the tidal air remains in the dead space; the rest reaches the alveoli, where it mixes with the air already in the alveoli, *i.e.*, the expiratory reserve volume plus the residual volume, approximately 3,000 cc. (Fig. 134). These volumes vary according to the individual and the conditions which obtain (rest, work, emotion, etc.), but on an average the volume of fresh air reaching the alveoli is approximately

[1] Unless otherwise specified the term "vital capacity" signifies Hutchinson's vital capacity as here defined. Other "vital capacities" are less frequently referred to.

[1] ALTAMIRA, J. DEL V., Estudios sobre la Capacidad Vital de los Estudiantes de Medicina de Córdoba, thesis for M. D., Buenos Aires, 1941.

10 to 12 per cent of the volume of air in the lungs at the end of expiration. In normal breathing at rest, therefore, one-tenth to one-eighth of the air in the lung is renewed at each inspiration. This proportion is known as the ventilation coefficient; it increases proportionately to the depth of breathing.

In certain circumstances it is useful to know the number of respiratory movements needed to renew all the air in the lungs. For instance in anesthesia with gases using the "closed-circuit" method, it is desirable to eliminate almost completely the nitrogen in the lung. The usual procedure is to start anesthesia with an "open circuit," *i.e.*, breathing the expired air out into the atmosphere, and to close the circuit only after about ten respiratory movements have been made. One method of knowing when a certain gas such as hydrogen has been completely eliminated is to have the subject inhale one breath of hydrogen and then to examine the hydrogen concentration in samples of expired air taken at each successive breath. It will thus be seen that in 6 to 10 respiratory movements almost all the hydrogen has been expelled from the lungs.

Ratio of pulmonary ventilation to oxygen absorption. This ratio, introduced by Anthony, refers to the liters of air breathed for each 100 cc. of oxygen absorbed by the lungs. The average values found in normal subjects are 2.75 liters (Anthony) or 2.44 liters (Knipping), the extremes being 1.68 and 4.5 liters. There are no considerable daily variations in the same subject; values found during exercise are similar to those found during rest; and the addition of oxygen to the inspired air does not alter the ratio. These observations show that, with adequate oxygen tension, CO_2 tension is the principal factor governing pulmonary ventilation.

The determination of this constant is of special value in cases of functional disturbance of the lung, particularly when the ventilation of each lung is considered separately (see "Bronchospirometry," below). It is more valuable as a sign of respiratory efficiency than the determination of the vital capacity. Normal pulmonary ventilation and vital capacity can be found in cases in which the ratio of pulmonary ventilation to oxygen absorption is high, indicating poor oxygen absorption, which is usually due to disturbances in the lungs or the pulmonary circulation.

Pneumography. This consists in taking a record of the respiratory movements (a pneumogram) by means of special instruments known as pneumographs. Usually the changes in the mid-thoracic perimeter are registered. The records show that expiration lasts longer than inspiration. If the respiratory rate is below 15 per minute, a short expiratory pause is recorded, but with higher frequencies this pause is no longer observed.

Respiratory frequency. The rate of respiration varies in normal subjects according to the species and is influenced by age, sex, size or height, work and rest, sleep, digestion, etc.

Table 22. Average Respiratory Frequency at Rest at Different Ages

Age	*Frequency per Minute*
Newborn	60–70
5 years	26
15–20 years	20
25–30 years	16

Respiratory frequency is a function of the metabolic rate. As the metabolic rate increases, so does the need of oxygen, and the rate of respiration also increases to supply this greater demand. In basal conditions animals of small size have a higher respiratory frequency than larger animals. The latter have a relatively smaller body surface with respect to weight, and a lower basal metabolic rate than the former. Frequency of respiration increases whenever the oxygen consumption increases; *e.g.*, during exercise, fever, etc.

Respiratory minute volume. The respiratory minute volume is obtained by multiplying the tidal volume by the respiratory frequency; it is 6 to 8 liters in a normal adult man when resting. This is the total pulmonary ventilation, but these figures do not always give a correct idea of its efficiency. For example, superficial respiration of great frequency does not ventilate the lungs adequately in spite of a large respiratory minute volume, because only a fraction of the tidal volume reaches the alveoli; to determine the alveolar ventilation, the dead space corresponding to each respiratory movement must be subtracted from the total ventilation.

Pulmonary ventilation, increased by a higher respiratory frequency and larger tidal volume, is also affected by a third factor, *i.e.*, an increase in respiratory surface,[1] which shows itself by an increase in total lung volume and in residual

[1] PEYSER, E., A. SASS-KORTSÁK, and F. VERZÁR, *Am. J. Physiol.*, **163**, 111, 1950; VERZÁR, F., *Ciencia e invest.*, 7, 303, 1951.

functional capacity. This has been demonstrated in several species, including man, in response to oxygen want caused either by physical exercise or by a fall in the oxygen partial pressure in inspired air. The greater lung volume is due, according to Verzár and his associates, to relaxation of the smooth muscles of the bronchi, brought about by sympathetic discharges, and to an increase in inspiratory muscle tonus. The larger alveolar surface facilitates diffusion of oxygen from the alveolar air to the blood. CO_2 in the inspired air, in spite of increased ventilation, had no effect on total lung volume.

Respiration is rhythmic when the respiratory movements are of the same depth and duration. Disturbances in respiratory rhythm will be considered in Chap. 33.

THE PHYSIOLOGY OF THE BRONCHI

The large and median bronchi are simply air tubes. The cartilage in their walls does not form complete rings, but it is sufficient to keep them open and to prevent any considerable change in their diameter. The small bronchi have no cartilage but have a layer of smooth muscle (Reisseissen's muscle). Contraction and relaxation of these muscle fibers modify the diameter of the bronchioles and thus the amount of air entering the alveoli, alveolar ventilation, and alveolar gas tensions. Moreover, a deep inspiration performed when the bronchioles are constricted causes alveolar pressure to fall considerably more than during normal inspiration, and thus may produce transudation of plasma from the pulmonary capillaries into the alveoli.

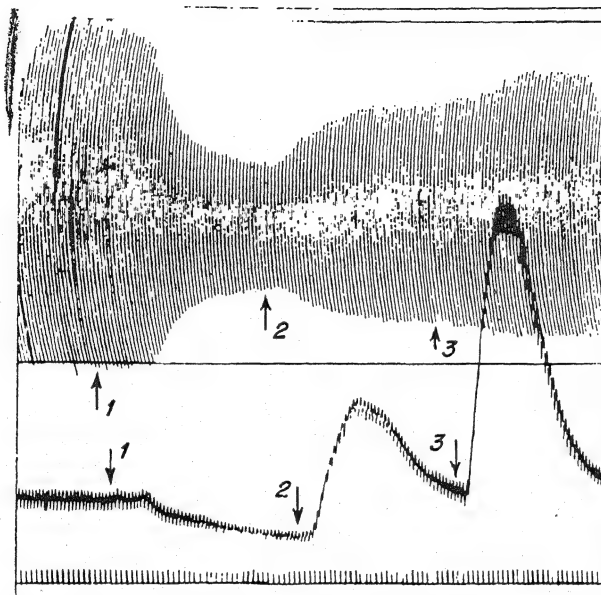



Fig. 135. Record of lung volume and blood pressure (upper and lower tracings respectively). 1, injection of morphine: the bronchioles contract and the lungs expand less at each respiratory movement; 2 and 3, adrenaline is injected and as the bronchioles relax more air enters the lungs and they expand more. Note the increase in blood pressure aftereach injection of adrenaline. Time in 4 sec. (*Jackson, D. E., "Experimental Pharmacology," Mosby, St. Louis, 1917.*)

Methods for studying bronchial musculature. Contraction of the bronchiolar muscles diminishes the amount of air flowing into the lungs; relaxation has the opposite effect. Therefore when the lungs are functioning rhythmically and at constant pressure, the volume of the lungs diminishes when the bronchiolar muscles contract and increases when they relax. The volume of the lung can be recorded by introducing into the pleural cavity a tube connected with a Marey tambour. At each insufflation the lungs expand. If a bronchioconstrictor substance is injected, expansion is considerably reduced (Fig. 135, 1); it can again be increased by injecting a bronchiodilator substance (Fig. 135, 2 and 3).

A variant of this method consists in measuring the pressure of the air in the lung which is being expanded rhythmically by a constant volume of air sent into the lung either by (*a*) insufflation into the trachea, (*b*)

compression of the thorax, or (*c*) compression of the lung placed in a plethysmograph.

In another method the isolated lung of a rabbit or guinea pig is perfused at constant pressure with Ringer's fluid, which is made to flow in through the trachea and bronchi and to flow out through minute openings made on the outer surface of the lung by scratching it with a needle. The amount of fluid passing through the lung in unit time depends on the diameter of the bronchioles. Bronchioconstrictor agents added to the perfusion fluid diminish or even stop the flow through the lungs.

Visceral motor nerves control the movements of the bronchial muscles. Stimulation of the parasympathetic contracts these muscles, and they are relaxed if the sympathetic is stimulated. Parasympathomimetic drugs (*e.g.*, acetylcholine) have a constrictor effect. Sympathomimetic drugs (*e.g.*, adrenaline) and parasympatholytic drugs (*e.g.*, atropine) relax the bronchiole muscles only if they are previously in a state of contraction; in normal conditions they have no effect. This observation shows there is no vagal tone, *i.e.*, the bronchiole muscles are not maintained in tonic contraction by vagal impulses, at least in experimental conditions. During an attack of asthma the bronchial muscles are contracted. Frequently it is possible to produce immediate relief by injecting a bronchiodilator agent (adrenaline, atropine, etc.).

EXPLORATION OF THE LUNG

The condition of the lungs can be examined by several methods. Percussion of the thorax over the parts of the lungs filled with air gives a resonant sound. On auscultation a peculiar rustling sound, the "vesicular murmur," is heard when air penetrates into the alveoli; it is an inspiratory sound, and increases in deep inspiration. It is not heard when air does not enter the alveoli. A murmur, due to the passage of air in the trachea and large bronchi, is heard if the ear is placed over the manubrium of the sternum or the stethoscope over the larynx. This murmur is normally not heard over the lateral walls of the thorax, but it can be heard in certain diseases of the lung, *e.g.*, pneumonia.

X-rays (radioscopy and radiography) are frequently used to determine the condition of the lungs and in the study of the mechanics of respiratory movements (Figs. 128, 129, and 132).

Bronchospirometry. The functions of each lung can be examined separately, thanks to the methods of bronchospirometry introduced by Jacobeus in 1932 and improved by Gebauer in 1939. The latter designed a catheter made of flexible rubber with two tubes. The longer of the two tubes is introduced into the left bronchus and adapted to it so that air can go into and come out of the left lung only through this tube. The shorter tube is adapted to the trachea, through which only air going and coming from the right lung circulates. The external ends of the two tubes are connected with two Benedict-Roth spirometers.

Oxygen consumption, respiratory minute volume, tidal volume, pulmonary ventilation, oxygen consumption ratio, vital capacity, inspiratory and expiratory reserve and residual volumes, and respiratory gaseous exchange for each lung can be determined by this method. This information is valuable to determine the respiratory deficiency in a diseased lung and the degree of compensation accomplished by the healthy lung, before applying certain therapeutic measures, such as artificial pneumothorax or the surgical removal of a lung or part of a lung.

In normal subjects lying on one side, the vital capacity of each lung varies according to the degree of compression to which it is submitted. If the curvature of the spine is such that the upper lung is compressed, the lung on the lower side has a greater vital capacity, respiratory minute volume, and inspiratory reserve volume and absorbs more oxygen.

The composition of alveolar air and the respiratory quotient of one lung can be different from that of the other lung in certain cases of pulmonary disease. The alveolar air collected by the usual method is a mixture of the alveolar air from both lungs.

THE UPPER RESPIRATORY TRACT

1. *Nasal cavities.* The alae nasi are dilated at each inspiration in some species (rabbit, horse). In man this occurs only when there is dyspnea.

The nasal mucosa plays an important part in modifying the inspired air as it passes through the nasal cavities, and thus it protects the more delicate membranes of the lung. The surface of the mucosa is increased by being spread out over the conchae. It is provided with a rich vascular network, and it secretes continuously. Small dust particles are retained by the secretions of the mucosa and by the nasal hairs. Evaporation of nasal secretions humidifies the inspired air. The temperature of the air is raised as it passes over the well-vascularized mucosae.

There are many nerve endings in the nasal mucosa. Those of the trigeminal nerve are stimulated by irritant vapors in the inspired air. A reflex is thus initiated which inhibits inspiration and provokes bradycardia. Section of the vagus nerve or injection of atropine prevents the reflex bradycardia. Olfactory nerve endings are situated in the upper part of the nasal mucosa; these will be considered in Chap. 75, when discussing the sensation of smell.

2. *Pharynx.* This is merely an air passage and has no other special function in respiration. It descends during inspiration, together with the larynx, being pulled down by the lungs, which expand downward as the diaphragm contracts.

3. *Larynx.* The glottis is dilated during inspiration by the separation of the vocal cords. When the vagus nerve or its recurrent laryngeal branch is paralyzed, the vocal cords remain flaccid and in accelerated breathing inspiration is disturbed. In young animals bilateral paralysis of the vocal cords causes death by asphyxia.

ARTIFICIAL RESPIRATION

When respiration ceases (drowning, carbon monoxide asphyxia, electrocution, etc.) or there is danger of respiratory failure, the air in the lungs must be renewed by artificial respiration. Many methods have been proposed, but only a few will be mentioned here.

Methods of artificial respiration can be divided into mechanical and manual. There are two types of mechanical methods, the so-called "resuscitators" and the iron lungs. A resuscitator[1] is an apparatus which produces ventilation of the lung by providing rhythmic variations of pressure through a face mask or tracheal catheter. Some completely mechanical models, *e.g.*, the spiropulsator, allow changes in respiratory rate and volume, length of inspiration, and length of inspiratory pause. In asphyxia of the newborn, oxygen (or 95 per cent oxygen + 5 per cent CO_2) is rhythmically insufflated through a soft rubber catheter inserted in the trachea; the pressure should not rise above a few centimeters of water in order not to damage the delicate lung tissues.

There are several models of "iron lung." Such an apparatus consists of a metal tank in which the patient is placed with the head protruding through a rubber collar. The pressure of the air in the tank is alternately increased and diminished by an electric pump. The alternating changes in pressure expand and compress the thorax, so that air is drawn into the lung and expelled from it. This type of apparatus has been used to keep patients alive for several months after the respiratory center has been damaged by disease (*e.g.*, poliomyelitis).

[1] COUNCIL ON PHYSICAL MEDICINE AND REHABILITATION, *J.A.M.A.*, **150**, 695, 1952.

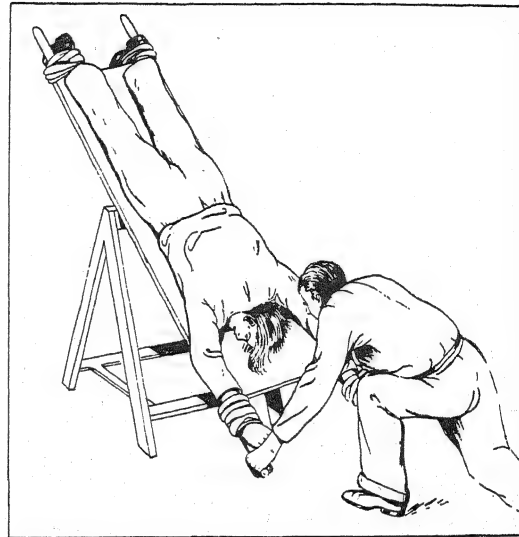



FIG. 136. Artificial respiration by the "rocking method." The abdominal viscera push the diaphragm into the thorax, and expiration is evoked when the subject is in the position illustrated; when the head is raised, the viscera pull on the diaphragm and cause inspiration. (*After Eve, F. C., J. A. M. A., vol. 124, p. 964, 1944.*)

In the *rocking method*,[1] the patient is placed in the prone position on a trestle that can be tilted so that either the head or the feet are down (Fig. 136). The patient is rocked by alternately tilting his head and feet to an angle of 45°, about ten times per minute. When the feet are down the abdominal viscera pull the diaphragm down and expand the lungs (inspiration). When the head is down the viscera push the diaphragm into the thorax (expiration). If the subject is a small child this method can be applied without the use of the trestle, by simply holding the child and rocking it.

Many manual methods of artificial respiration have been proposed, but until recently it was difficult to judge their efficiency comparatively.

[1] EVE, F. C., *J. A. M. A.*, **124**, 964, 1944.

In order to do so, Gordon and his associates[1] experimented on 26 volunteers who were anesthetized and rendered completely apneustic by an adequate dose of curare. The tidal volume was measured and compared using six methods, changing the position of the subject and the technique employed.

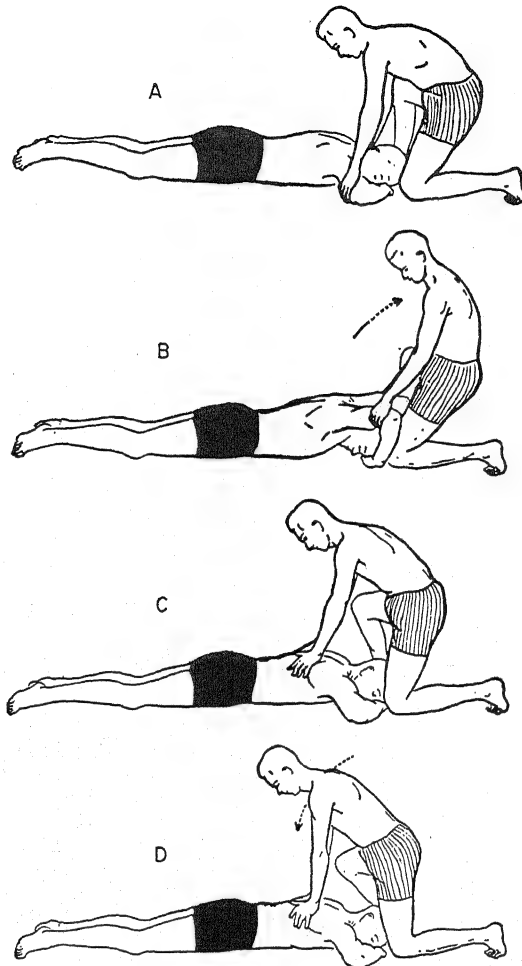



FIG. 137. Back pressure-arm lift method (after Holger-Nielsen). *A*, placing hands for arm lift; *B*, arm lift; *C*, placing hands for back pressure; *D*, back pressure. (*Gordon, A S., et al., J. A. M. A., vol. 147, p. 1444, 1951.*)

A manual method of artificial respiration should be simple, so that it can be easily learned by anybody. It should not tire the operator, so that it can be kept up for a long time. Methods in which the operator provokes both respiratory movements are superior to those in which only inspiration or expiration is produced. Schäfer's prone pressure method is purely expiratory; the thorax is compressed to expel air, and inspiration is due to the subsequent elastic expansion of the thorax. In Gordon's experiment, it was not found satisfactory; the mean tidal volume was only 485 cc., and in some of the subiects cyanosis could not be abolished.

The method recommended by the Council on Physical Medicine and Rehabilitation of the American Medical Association[1] is the so-called "back pressure-arm lift method" (Fig. 137).

"The subject is placed prone, elbows bent, arms overhead with one hand upon the other. The cheek is placed on the hand, the face turned slightly to one side. The operator kneels on one knee at the head of the subject and puts the foot of the opposite leg near the elbow. He places his hands on the subject's back, in such a way that the thumbs just touch, the heels of the hands being just below a line running between the armpits. He rocks slowly forward, elbows straight, until his arms are approximately vertical, exerting steady pressure upon the chest. Then he commences to rock backward slowly and slides his hands to the subject's arms just above the elbows. He raises the arms until resistance and tension are felt at the subject's shoulders. Then he drops the arms. This completes a full cycle. The cycles are repeated 12 times per minute, the expansion and compression phases being of equal length, the release periods being of minimum duration." Both respiratory movements are thus actively produced. The mean tidal volume was found to be over 1,000 cc. The prone position of the subject has the advantage that if the tongue has lost its tonus it does not fall back obstructing the upper respiratory path. Moreover, if the patient vomits, it will be less likely that foreign bodies are drawn into the trachea.

Resuscitation often requires not only artificial respiration but also the administration of oxygen and CO_2. Circulatory conditions should be taken care of, especially in circulatory collapse, and the body temperature should be maintained.

[1] GORDON, A. S., *et al., J. A. M. A.*, **147**, 1444, 1951.

BIBLIOGRAPHY

ANTHONY, A. J., "Funktionsprüfung der Atmung," Leipzig, 1937.

BJORKMAN, Bronchospirometrie, *Acta med. Scandinav.*, Stockholm, Suppl. LVI, 1934.

[1] COUNCIL ON PHYSICAL MEDICINE AND REHABILITATION, *J. A. M. A.*, **147**, 1454, 1951.

FELIX, W., Anatomie der Atmungsorgane, in Bethe's "Handbuch der normalen und pathologischen Physiologie," Springer, Berlin, **2**, 37, 1925.

FENN, W. O., Mechanics of Respiration, *Am. J. Med.*, **10**, 77, 1951.

GORDON, A. S., M. S. SADOVE, F. RAYMON, and A. C. IVY, Critical Study of Manual Artificial Respiration, *J. A. M. A.*, **147**, 1444, 1951.

ROGER, G. H., and L. BINET, Respiration, in "Traité de physiologie normale et pathologique," vol. 5, Masson et Cie, Paris, 1934.

ROHRER, F., Physiologie der Atmenbewegung, in Bethe's "Handbuch der normalen und pathologischen Physiologie," Springer, Berlin, **2**, 70, 1925.

ROSS, B. D., Five-year Survey of Methods for Artificial Respiration, *J. A. M. A.*, **129**, 443, 1945.

VACCAREZZA, O. A., "Los Músculos Respiratorios en la Fisiopatología Pulmonar," Buenos Aires, 1940.

CHAPTER 27

Exchange of Gases in the Lung

THE EXCHANGE OF gases through the alveolar wall between the air and blood is known as hematosis. The total surface of the alveoli has been calculated to be between 80 and 100 sq. m. in man. Their walls are sufficiently thin and permeable to allow of a rapid interchange of gases. In mammals the interchange of gases takes place between a gaseous environment and a liquid *milieu intérieur*. The fundamental laws of gases must therefore be known in order to understand the phenomena of hematosis.

THE GAS LAWS

The laws governing the gases taking part in hematosis are the laws common to all gases; they apply not only to oxygen and carbon dioxide, but also to inert gases (nitrogen, hydrogen, etc.), toxic gases and vapors (carbon monoxide, hydrocyanic acid, hydrogen sulfide, benzene vapor, etc.), and those used in therapeutics (anesthetics, amyl nitrite, etc.). Through the alveolar walls, water is eliminated in the form of aqueous vapor in the expired air. Heat is also lost through the respiratory surface. Solid particles (silica, coal dust, etc.) are absorbed through the surface.

The principal laws governing the interchange and transport in solution of gases will be briefly summarized here; further details will be found in textbooks on physics.

The Boyle-Mariotte law. Experiments performed by Boyle and Mariotte showed that when the temperature remains constant the pressure of a gas varies inversely with its volume, or conversely the volume varies inversely with the pressure. If p_1 and p_2 are the pressures corresponding to volumes v_1 and v_2,

$$\frac{p_1}{p_2} = \frac{v_2}{v_1}$$

therefore $\quad p_1v_1 = p_2v_2$

whence $\quad p_nv_n = K$

The Boyle-Mariotte law can be stated as follows: at a constant temperature the product of the pressure and volume of a gas has a constant value.

The law of Gay-Lussac (or of Charles). The second fundamental law of gases, discovered by Gay-Lussac, refers to the influence of temperature. At a constant pressure the volume of a gas increases with the temperature. For each rise in temperature of 1°C. the gas expands $1/273$ of its volume at 0°C. This fraction $1/273$ is equal to 0.003663, which is known as the dilatation coefficient of gases. Gay-Lussac's law can be expressed as follows:

$$\frac{V_1}{V_2} = \frac{T_1}{T_2}$$

the volume of gas being directly proportionate to absolute temperature.

If the volume is maintained constant when the gas is heated, the pressure rises $1/273$ of the pressure at 0°C. for each degree centigrade that the temperature is raised; this is known as the pressure coefficient of gases. In this law the interdependence of pressure and volume is again evident, since the volume and pressure coefficients of gases are the same.

Practical applications. The dilatation coefficient of gases is much greater than that of solids and liquids. On the other hand a gas also has great compressibility. For example, at a constant pressure (*e.g.*, 760 mm. Hg) 1 liter of air at 0°C. dilates to 1,073 cc. at 20°C., an increase of 7.3 per cent above the initial volume. These differences are not of great importance when measuring gases with only a rough approximation, as when the vital capacity is measured in a spirometer; but when exact measurements must

be taken (*e.g.*, when the alkaline reserve is measured in volumes of CO^2 per cent, or when the oxygen content of blood is determined) the exact temperature and pressure at which the volume has been measured must be known. Without these data the volume of a gas has no precise significance. To avoid confusion in the exact measurement of a gas, the volume is given

Table 23. Barometric Pressure at Different Heights above Sea Level

Height above sea level		*Barometric pressure, mm. Hg*	*Temperature of boiling water, °C.*	*Oxygen partial pressure, mm. Hg**	*Equivalent O_2, %†*
M.	*Ft.*				
20,000	65,600	44	35	9.2	1.21
18,000	59,000	60	41	12.5	1.65
16,000	52,500	82	47	17	2.26
14,000	45,950	111	53	23	3.30
12,000	39,370	151	60	31	4.15
10,000	32,800	205	67	43	5.65
8,000	26,250	274	73	57	7.55
7,000	23,000	315	77	66	8.66
6,000	19,685	360	80	75	9.8
5,000	16,400	411	83	86	11.3
4,000	13,120	466	86	98	12.8
3,000	9,845	529	90	110	14.5
2,000	6,560	598	93	125	16.5
1,000	3,280	675	96	141	18.6
0	0	760	100	159	20.94

* Calculated for dry air.
† Calculated for a total pressure of 760 mm. Hg.

as if it were made at 0°C. and at 760 mm. Hg pressure. The volume read is multiplied by $(273/273 + t) \times (P/760)$, t being the temperature in degrees centigrade and P the barometric pressure in millimeters of mercury.

Pressure measurements and units. The air surrounding the earth is subject to the force of gravity and at sea level exerts a pressure equal to that of a column of mercury 760 mm. high. This is known as a pressure of 1 atmosphere and is approximately equal to 1 kg. per sq. cm. (14.7 lb. per sq. in.). A column of mercury of 1 sq. cm. cross section and 76 cm. height is made up of 76 cc. of mercury, which weighs $76 \times 13.5951 = 1{,}033.2$ gm. (13.5951 is the specific gravity of mercury). This column exerts on its base of 1 sq. cm. a pressure of 1,033.2 gm., and is equivalent to the pressure exerted by the atmosphere on each square centimeter at sea level. When, therefore, it is said that a pressure of 1 atmosphere is equivalent to 1 kg. per sq. cm., an error of 3 per cent is committed, but this error is of no importance in the usual measurements.

A pressure of 1 atmosphere is also equivalent to that of a column of water of 10.33 m. height. Therefore the pressure to which a diver is submitted increases by 1 atmosphere for approximately every 10 m. depth.

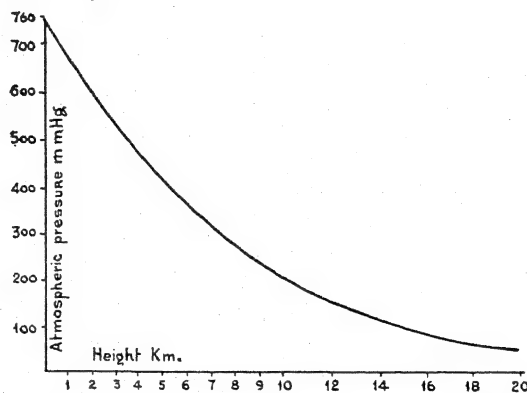



FIG. 138. Barometric pressure at different heights above sea level. Small differences due to changes in temperature are not taken into account in the diagram.

Atmospheric pressure. When a person ascends in an airplane or goes up a mountain, the column of air above him decreases as he rises, and the pressure also decreases (Table 23, Fig. 138).

Solubility of a gas in a liquid. If a gas and a liquid are placed in a closed flask, the gas dissolves in the liquid until an equilibrium is reached. If the gas does not combine with the liquid but is simply dissolved in it, the amount of gas absorbed by the liquid depends on (*a*) the nature of the gas and the liquid; (*b*) the temperature; (*c*) the partial pressure of the gas.

The volume of a given gas absorbed by a liquid at a given temperature is constant whatever the pressure. For example, a liquid is placed in a dish within a cylinder closed by a piston (Fig. 139); it dissolves 2 cc. of gas when the piston is in position *A*. The volume of gas in the cylinder is then reduced to one-half by placing the piston in position *B*, and time is allowed for an equilibrium to be established between the gas and the liquid. The amount of gas absorbed is still 2 cc., but at twice the pressure. Therefore at normal temperature and pressure (NTP), *i.e.*, at 0°C. and 760 mm. of mercury, the gas dissolved by the liquid when the piston is in position *B* will have twice the volume of the gas absorbed when the piston is in position *A*.

When the temperature is maintained constant, the amount of gas dissolved by a liquid is direct function

of pressure; in a system of cartesian coordinates this is represented by a straight line (Fig. 145, solution of CO_2). Henry's law of solution of gases states that, if the temperature is maintained constant, the volume of gas, measured at 0°C. and 760 mm. Hg, in solution in a liquid is proportional to the pressure of the gas.

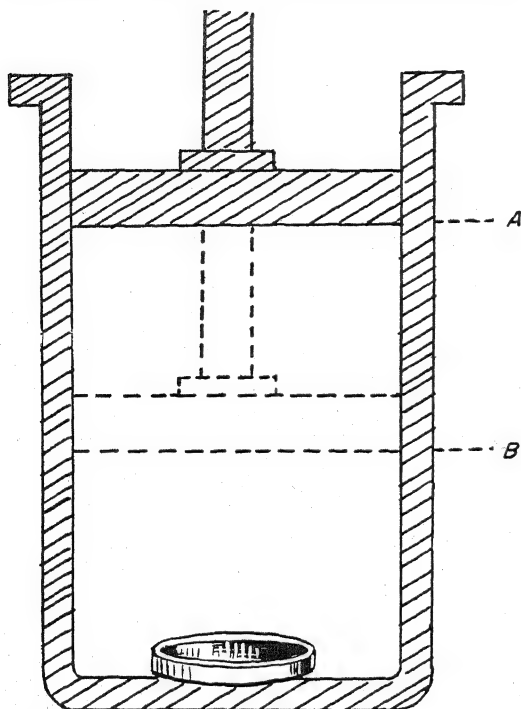



FIG. 139. Diagram of apparatus for demonstrating the effect of pressure on the solution of gases in liquids. When the piston is in position *B*, twice as great a volume of gas is dissolved by the liquid as is dissolved when the piston is in position *A*.

The absorption or solution coefficient is the volume of gas (corrected for NTP) absorbed by 1 cc. of the liquid at a pressure of 760 mm. Hg. The absorption coefficient diminishes as the temperature increases. Substances in solution in a liquid diminish its capacity to dissolve gases.

Dalton's law of partial pressure. In gas mixtures the pressure exerted by each gas depends on the percentage of the gas in the mixture. The pressure exerted by each gas in the mixture is known as the partial pressure of the gas. For example, in a gas mixture of 20 per cent oxygen and 80 per cent nitrogen at a pressure of 760 mm. Hg, the partial pressure of oxygen is $(760 \times 20)/100 = 152$ mm. Hg, and that of nitrogen $(760 \times 80)/100 = 608$ mm. Hg. The sum of the partial pressures is the total pressure, *i.e.*, $152 + 608 = 760$ mm. Hg. Therefore

$$\text{Partial pressure} = \frac{\text{Total pressure} \times \text{Percentage concentration of gas}}{100}$$

For example, if the concentration of CO_2 in alveolar air is 5.5 per cent, and the atmospheric (total) pressure is 760, the partial pressure of CO_2 is $(760 \times 5.5)/100 = 41.8$ mm. Hg.

When calculating the partial pressure of a gas, all the gases in the mixture must be taken into account, including the tension of water vapor. This is particularly important when calculating the partial pressure of gases in alveolar air, which is saturated with water; water vapor at 37°C. has a tension (partial pressure) of 47 mm. Hg. (Table 24). For example, when calculating the partial pressure of CO_2 and O_2 in a sample of alveolar air which has 5.3 per cent CO_2 and 14 per cent O_2, the barometric pressure being 756, first the tension of water vapor (47 mm. Hg at 37°C.) must be subtracted from the total pressure: $756 - 47 = 709$. The partial pressure of CO_2 is 5.3 per cent of 709, *i.e.*, 37.6 mm. Hg, and that of O_2 is 14 per cent of 709, *i.e.*, 99.3 mm. Hg.

Table 24. Tension of Water Vapor at Different Temperatures

Temperature, °C.	*Tension of Water Vapor, Mm. Hg*
0	4.6
10	9.2
20	17.5
30	31.8
37	47.0
40	55.3
50	92.5

An understanding of partial pressure is important when a mixture of gases is in contact with a liquid, because each gas goes into solution independently of the others, in an amount directly proportional to its own partial pressure (Dalton's law). The fixation of oxygen by hemoglobin is also dependent on the oxygen partial pressure. It is therefore important to know fairly accurately the minimum partial pressures of oxygen compatible with the normal functioning of the tissues. Recent developments in aviation, pathology, etc., have enhanced the value of this knowledge.

Equilibrium between a liquid and a gas. If a certain amount of liquid is placed in a closed flask (Fig. 149) and the flask is then filled with a gas mixture of known composition, an interchange of molecules between the liquid and the gas will commence and, according to Henry's and Dalton's laws, will continue until an equilibrium is established. Once the equilibrium is established the amount of

gas in solution remains constant, because for each molecule of gas that goes into solution, another is set free. The gases are at the same tension in the liquid as in the gas. The tension of a gas in a liquid is determined by the method known as tonometry. The time taken for the establishment of this equilibrium depends on the area of the surface of contact in relation to the volume of liquid. To increase the surface of contact, the flask is placed horizontally and rotated.

Combination of gases with liquids. Henry's law is valid only when a gas goes into solution in a liquid; if there is chemical affinity between the gas and the liquid, or a substance dissolved in the liquid, either of two possibilities may occur:

1. Chemical combination may not be dependent on the partial pressure of the gas; *e.g.*, CO_2 in contact with a solution of Na(OH) is absorbed whatever the partial pressure as long as there is Na(OH) in solution.
2. Chemical combination may be dependent on partial pressure The amount of gas fixed by the solution or dissolved substance then decreases if its partial pressure is lowered, and vice versa, but not necessarily in direct proportion; *i.e.*, Henry's law is not applicable. The gas combined is set free in a vacuum. The combination of oxygen with hemoglobin corresponds to this second type.

Diffusion. Subs ances in solution migrate from the more to the less concentrated parts of the solvent. In the same way gases diffuse from where the partial pressure is higher toward where it is lower, at a speed that increases with the difference in pressure.

Krogh's diffusion constant is the volume of a gas (in cubic centimeters) that passes through the pulmonary membranes in 1 min., when the difference in the partial pressure of the gas in the alveolar air and the blood in the pulmonary capillaries is 1 mm. Hg. The oxygen diffusion constant varies from a minimum of 20 in the resting condition to a maximum of 60 in exercise.

METHODS OF GAS ANALYSIS

The methods used in the study of the respiratory function of the blood (gas transport) and hematosis are of interest in physiology, applied physiology (*e.g.*, aviation physiology), and pathology. Only the more current methods will be briefly considered; further details can be obtained in technical manuals.

Determination of CO_2 and O_2 in air. One of the modifications of Haldane's buret is most commonly used for this purpose. A sample of air is collected and its volume accurately measured. The air is then washed in a flask containing KOH, which fixes CO_2. The volume of air is again measured; the difference found corresponds to the CO_2. Oxygen is then fixed by an alkaline solution of pyrogalate and the volume is again measured; the difference corresponds to the O_2. The remainder consists of inert gases: nitrogen, helium, argon, etc. Other gases (CO, anesthetic gases, etc.) can be measured in this apparatus appropriately modified. The CO_2 and O_2 content of the atmosphere, expired air, alveolar air, etc., can be determined by this method. Alveolar air is collected by making a forced expiration into a long tube and then closing it with the tip of the tongue. A lateral tube placed near the extremity close to the mouth is connected with a sampling bulb into which a sample of the last air expired is drawn.

Determination of the gases in the blood. Van Slyke and Neill's[1] apparatus is commonly used for this purpose. A sample of liquid is taken and the gases it contains are set free by simple vacuum evacuation or by means of chemical substances; the volume is measured by an indirect method and corrected for NTP. The percentage of the gas in the mixture is determined by this procedure, but not its partial pressure; to obtain the latter a tonometric method must be used.

Tonometry. To determine the partial pressure of a gas in blood (*e.g.*, oxygen in renal venous blood), a sample is taken by puncturing the vein and collecting the blood under mineral oil in a tube containing an anticoagulant agent. The content of oxygen of a given volume is measured; let us suppose a concentration of 12 per cent is found. Small amounts of this blood are then distributed in a series of flasks (Fig. 149), which are then filled with mixtures of gases with varying oxygen concentration (10, 11, 12 per cent, etc.). The flasks are closed and gently rotated until a gaseous equilibrium is established. The oxygen content of the blood in each flask is then determined. It will be found to be below 12 per cent in some and above 12 per cent in others, but in one of the flasks it will remain constant at 12 per cent. The blood taken from the vein therefore was in gaseous equilibrium with the gas in that flask (let us suppose it contained 15 per cent oxygen). The barometric pressure and temperature are read, the tension of water vapor is subtracted, and the oxygen partial pressure of the gas mixture is calculated. This oxygen partial pressure is the same as that of the blood sample

[1] VAN SLYKE, D. D., *et al.*, *J. Biol. Chem.*, **61**, 523, 1924.

fishes, in which the air has a high O_2 concentration. The O_2 partial pressure in the swim bladder is much higher than in the tissues; therefore it cannot have accumulated there by simple diffusion, but must have been actively secreted. The epithelial lining of the swim bladder has a glandular aspect and is controlled by the nervous system, conditions that are not similar to those of the alveoli. The function of the swim bladder differs considerably from that of the lungs, and it is not justifiable to apply facts observed in the former to events occurring in the latter.

Physical laws alone are sufficient to explain the process of diffusion of gases through the alveolar capillary membrane. Experiments by Hartridge and Barcroft have shown that the O_2 partial pressure in the alveolar air is always slightly above that in the blood. It has also been calculated that 250 cc. per min. of O_2 can diffuse through the alveolar membrane for every millimeter of difference in the O_2 partial pressures on each side of the membrane.

Normally the difference in O_2 partial pressure between the alveolar air and venous blood is approximately 60 mm. Hg and the oxygen consumption at rest is 300 cc. per min. The oxygen consumption can rise to 4,000 cc. per min. in maximum muscular effort. In a normal individual, therefore, the absorption of O_2 is not limited by its capacity to diffuse through the lungs but by other factors, *i.e.*, pulmonary ventilation (the speed at which the alveolar air is renewed) and the circulatory factor (the speed at which the erythrocytes circulate through the lung capillaries). The oxygen capacity of hemoglobin is constant; approximately 95 per cent is saturated at the O_2 partial pressure that exists at sea level. At high altitude another factor comes into play which limits oxygen absorption, *i.e.*, the decrease in O_2 partial pressure (see "Anoxia," Chap. 33).

When the organism is producing a maximum muscular effort the factor that limits oxygen absorption is the maximum amount of blood that can circulate through the lungs in unit time.

CO_2 has a speed of diffusion approximately 25 times that of O_2. It has been calculated that a difference in CO_2 tension of 0.03 mm. Hg between the blood and alveolar air would be enough to assure the elimination by simple diffusion of all the CO_2 produced by a normal person at rest. In the elimination of CO_2 an important part is played by carbonic anhydrase (see Chap. 30).

In the interchange of gases in the lung, the thickness of the membrane (alveolar and capillary) that separates the blood and the alveolar air is also important in pathologic conditions but not in normal ones. The presence of secretions or frothy edematous fluid in the alveoli, or of edematous thickening of the membrane, represents an obstacle to the diffusion of O_2. The elimination of CO_2 is not hindered so much because of its greater diffusion velocity.

Absorption of gases from closed cavities. The air in a closed cavity within the organism is gradually absorbed. Thus, when a bronchus is obstructed the air in the corresponding part of the lung is reabsorbed and atelectasis follows. The air in a pneumothorax is also reabsorbed. The mechanism by which air is reabsorbed is as follows:[1] The sum of the partial pressures of gases in arterial blood is the same as the atmospheric pressure, but in venous blood it is 54 mm. Hg less, because the considerable decrease in O_2 partial pressure caused by the absorption of O_2 by the tissues is not counterbalanced by the slight increase in CO_2 partial pressure (see Table 25). This causes the gases in a closed cavity to diffuse out into the surrounding tissues according to their velocity of diffusion and the capacity of the blood to dissolve them. After a few days the air in the cavity is stabilized and has the same O_2 and CO_2 content as the tissues. Reabsorption of the gaseous mixture is delayed by the presence of N_2, which is less soluble than O_2 and CO_2. The cavity is adjusted to atmospheric pressure and is gradually reduced until it disappears.

BIBLIOGRAPHY

See references at the end of Chap. 30.

[1] HENDERSON, Y., and M. C. HENDERSON, The Absorption of Gas from Any Closed Space within the Body, *Arch. Int. Med.*, **49**, 88, 1932.

CHAPTER 28

The Transportation of Oxygen by the Blood

OXYGEN TENSION IN the alveolar air is approximately 100 mm. Hg; at this tension arterial blood contains 19 volumes per cent of oxygen. The solubility coefficient of oxygen in blood at a pressure of 760 mm. Hg and at 38°C. is 0.022; at 100 mm. and at the same temperature, blood can carry only 0.3 volumes per cent oxygen in solution. The greater part of the oxygen in the blood is in unstable combination with hemoglobin.

The molecular weight of hemoglobin has been found to be between 63,000 and 66,400 (see Chemical properties of hemoglobin, Chap. 4) and each molecule has four atoms of iron. To simplify calculations it is usual to divide the molecular weight by four, so that 16,400 gm. hemoglobin corresponds to 1 gram atom of iron (56 gm.). This amount of hemoglobin combines with 1 gram molecule of oxygen at 760 mm. Hg and at 0°C, *i.e.*, with 22.41 liters. One gram of hemoglobin, therefore, combines with 1.34 cc. O_2, a fact that has been established repeatedly by experimental determination. Normal blood has approximately 15 gm. per cent hemoglobin, and this would hold

$$1.34 \times 15 = 20.1 \text{ cc. } O_2$$

if the hemoglobin were completely saturated. At an O_2 tension of 100 mm. Hg, hemoglobin is only 95 per cent saturated; therefore arterial blood with a hemoglobin content of 15 gm. per cent holds 19.1 cc. O_2. Thus it is possible to calculate the hemoglobin concentration in blood by determining its O_2 content at a given O_2 tension, or vice versa the oxygen content if the hemoglobin concentration and the O_2 tension are known.

Hemoglobin enters into chemical combination with O_2, since both substances are in definite and constant stoichiometric proportions, and the formation heat of HbO_2 is also constant.

The combination of Hb and O_2 is reversible, *i.e.* $Hb + O_2 \leftrightarrows HbO_2$. The direction in which it takes place depends on the following factors: oxygen partial pressure, temperature, pH, and the electrolytes in the solution. This unstable combination is usually known as "oxygenation" and not "oxidation" of hemoglobin.

Oxygen dissociation curve of hemoglobin. Conditions governing the fixation of oxygen by hemoglobin have been determined experimentally. In order to do this a few cubic centimeters of pure hemoglobin solution at a known concentration are placed in appropriate containers. The containers are then filled with gas mixtures of O_2 and N_2, varying the oxygen partial pressure by 10 mm. Hg from 0 to 100 mm. Hg. The blood is brought into gaseous equilibrium with the gas mixture, and the amount of oxygen absorbed by the blood at different O_2 tensions is determined, establishing the percentage of O_2 saturation, *i.e.*, the proportion of O_2 absorbed in relation to maximum absorption (1.34 cc. O_2 per gm. hemoglobin). The data obtained are plotted on *x-y* coordinates. The resulting curve is a hyperbole and is known as the "oxygen dissociation curve of hemoglobin, as a function of O_2 partial pressure" (Fig. 141).

If the experiment is repeated with blood instead of a solution of pure hemoglobin, a curve of different shape is obtained. At one time it was

thought that crystallized hemoglobin differed from hemoglobin in blood, but later it was established that the difference in the dissociation curves was due to the electrolytes in the blood.

The principal factors that modify the O_2 dissociation curve of hemoglobin are (*a*) temperature; (*b*) electrolytes; (*c*) hydrogen ion concentration (pH).

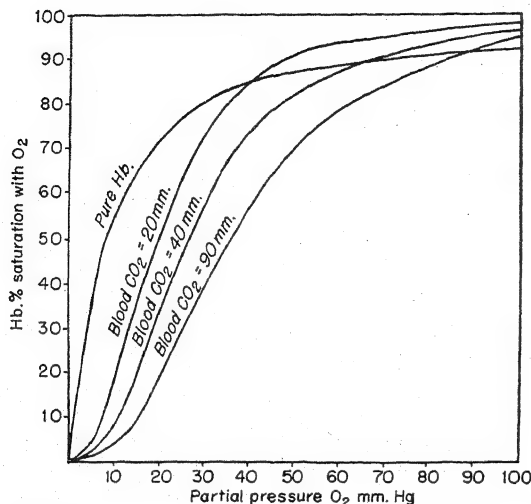



FIG. 141. Dissociation curve of hemoglobin at different oxygen pressures. The curve corresponding to a solution of pure hemoglobin is a regular hyperbole. Those corresponding to blood are irregular because of the electrolytes in solution. Presence of CO_2 shifts the dissociation curve to the right; the curve of normal blood is that corresponding to a CO_2 partial pressure of 40 mm. Hg.

An increase in temperature diminishes the capacity of O_2 fixation, *i.e.*, favors the dissociation of HbO_2. During tissue activity (*e.g.*, muscular contraction) there is a local rise in temperature. This favors release of O_2 from hemoglobin and its delivery to the tissues, in circumstances in which the need of O_2 increases.

Electrolytes, *i.e.*, the different ions in the blood plasma and erythrocytes, modify the O_2 dissociation curve of hemoglobin. The effect varies with the ion and its concentration. Pure hemoglobins prepared from blood of different mammalian species have approximately equal oxygen capacity, but the O_2 dissociation curve of the whole blood varies from one species to another. This is due to the electrolytes in the blood, as is shown by the fact that, although the O_2 dissociation curves of human and dog's blood are not the same, nevertheless a solution of pure human hemoglobin has an O_2 dissociation curve similar to that of dog's blood if electrolytes are added to the solution in the same concentration as they are found in dog's blood.

An increase in hydrogen ion concentration diminishes the oxygen fixation capacity of the blood, and a decrease increases the O_2 capacity (Fig. 141). Thus CO_2 entering into the capillaries from the tissues acidifies the blood and releases oxygen from hemoglobin, while in the lungs the loss of CO_2 into the alveolar air diminishes the acidity of the blood and increases its oxygen fixation capacity. Other conditions in which hydrogen ion concentration is of importance will be considered in Chap. 33.

The interaction of the different factors results in release or fixation of O_2 by hemoglobin according to the needs of the organism. Several examples of interdependent equilibriums will be given when dealing with the transportation of CO_2 by the blood.

Passage of oxygen into the tissues. Gases in arterial blood remain without change until the blood arrives at the capillaries. The thin capillary walls are in direct contact with the tissue cells or fluids. The oxygen partial pressure in the tissues is low (30 mm. Hg or less). Oxygen therefore diffuses from the erythrocytes to the plasma, and through the capillary wall into the tissue fluids and the cells, owing to the descending gradient in partial pressure. Moreover, dissociation of oxygen from hemoglobin is facilitated by the increase in acidity (CO_2 and sometimes lactic acid) and, when there is great activity, by an increase in local temperature. The amount of oxygen that blood releases to the tissues is conditioned by the activity of the tissues and the velocity of the circulation.

Oxygen reserve. The average oxygen saturation of arterial blood is 95 per cent. The breathing of pure oxygen has little effect on this percentage. Arterial blood holds 19 volumes per cent O_2, and venous blood holds 14 volumes per cent. The total blood volume in a normal adult is approximately 5 liters; if half of this is arterial blood and the other half is venous, the total amount of O_2 in the blood is 850 cc. To this must be added a small quantity of O_2 dissolved in the tissues, and the O_2 in alveolar air, to obtain the total O_2 in the body. Oxygen consumption at rest is approximately 300 cc. per min., and considerably more during exercise. Therefore if the gaseous interchange in the lungs

is stopped, there is scarcely enough O_2 in the body to supply its needs during 3 min. when resting, and an even shorter time when working. A few minutes after the gaseous interchange in the lungs has been stopped, therefore, symptoms of asphyxia appear. Anoxia (lack of oxygen) causes severe and even fatal disturbances if the respiratory interchange is not renewed without delay. Mammals, including man, have no oxygen reserve, but poikilothermic animals can live for some time without taking in oxygen, because they have a much lower rate of oxygen consumption.

BIBLIOGRAPHY

See references at the end of Chap. 30.

CHAPTER 29

Physicochemical Equilibriums in the Blood and Tissues

THE EXISTENCE IN the blood of many interdependent equilibriums has been well established. For instance, a change in CO_2 concentration modifies the pH, alkali reserve, Hb-HbO_2 ratio, cell Cl–plasma Cl ratio, etc. The transportation of gases is closely dependent on these important physiologic equilibriums. To have a clear idea of their significance certain fundamental physicochemical concepts, such as hydrogen ion concentration (pH), buffer systems, etc., must be well understood.

Concentration. The concentration of a substance is usually given by the weight or volume of the substance contained in the total volume in which it is diluted. Thus a 10 per cent solution of glucose is one that contains 10 gm. glucose in 100 cc. of the solution. In certain cases this way of expressing concentration is quite clear, *e.g.*, 0.9 per cent NaCl, or 2 per cent $CuSO_2$ need no further explanation. In other cases further data must be given to avoid confusion; thus a 40 per cent solution of alcohol may contain 40 cc. alcohol in 100 cc. of solution (volume to volume ratio, V/V), or 40 gm. alcohol in 100 cc. of solution (weight to volume ratio, W/V).

The physicochemical concept of concentration is somewhat different, and it is always given in gram molecules (or gram atoms, or gram ions) in unit volume of 1 liter. Physicochemical concentration is usually expressed by the letter c preceding the chemical symbol for the substance, or by including the symbol in brackets; thus cCa, or [Ca] signifies calcium concentration. For example, the molecular weight of glucose is 180 gm.;[1] therefore a solution containing 180 gm. per liter (W/V) in physical chemistry is expressed thus: [glucose] = 1 or 1 mole, because this solution contains 1 gram molecule per liter. A solution containing 1.8 gm. of glucose per liter (*i.e.*, $\frac{1}{100}$ gram molecule) is expressed as [glucose] = 0.01 or 0.01 mole, etc.

The atomic weight of Na is 23 and that of Cl is 35.5; therefore the molecular weight of NaCl is 58.5. A solution containing 5.85 gm. of NaCl ($\frac{1}{10}$ gram molecule) in 1 liter is expressed as [NaCl] = 0.1. This solution contains 0.1 gram atom of Na and 0.1 gram atom of Cl; therefore the concentrations of these atoms are expressed as [Na] = 0.1 and [Cl] = 0.1. If 90 per cent of the NaCl is dissociated into its constituent ions Na^+ and Cl^-, the concentration of Cl^- ion will be 90 per cent of 0.1, *i.e.* 0.09, when dealing with physicochemical concepts, and it will be expressed as $[Cl^-] = 0.09$. In weight of Cl this means there is $35.5 \times 0.09 = 3.195$ gm. of Cl^- ion per liter of solution.

This apparently more complicated way of expressing concentrations has many advantages and is really simpler in certain cases; the student will therefore find it useful to become familiar with this terminology. Thus two substances that are in the same concentration, in physicochemical parlance, have an equal number of molecules, or atoms, or ions in solution. For example, by chemical analysis it has been found that normal human blood contains, on an average, 250 mg. per cent of Cl and 160 mg. per cent of Na. These figures do not tell immediately what is the ratio of Na to Cl in blood, but they

[1] It is exactly 180.0936, because the atomic weight of H is not 1, but 1.0078. Where greater accuracy is not needed, molecular weights are usually given with only the first decimal.

can be converted to moles or millimoles (one thousandth part of a mole). One mole of Na weighs 23 gm. and one millimole 23 mg.; therefore 160 mg. Na per cent equals 6.96 millimoles per cent or 69.6 millimoles per liter (0.0696 mole). In the same way, 250 mg. per cent Cl equals 70.4 millimoles per cent and 0.0704 mole per liter. The concentration of Na and Cl in blood is expressed in physicochemical terminology as: [Na] = 0.0696 and [Cl] = 0.0704, and immediately it will be seen that total blood contains more atoms of Cl than of Na.

Another example will emphasize the advantage of using the physicochemical way of expressing concentration. The molecular weight of hemoglobin is 16,400 gm. per atom of iron. One mole of hemoglobin fixes one mole of molecular oxygen (O_2), *i.e.*, 22.41 liters. Therefore blood containing 16.4 gm. per cent of hemoglobin has 1 millimole per cent of hemoglobin, which can take up 1 millimole of O_2, *i.e.*, 22.41 cc. of oxygen for every 100 cc.

When the concepts of pH and buffer systems are dealt with, it will be seen how important it is to be familiar with this terminology.

Degree of dissociation. An electrolyte dissolved in water is dissociated into positive (+) ions, or cations, and negative (−) ions, or anions. Dissociation progresses until a dynamic equilibrium is established in which for every molecule dissociated another is formed by the union of the two ions in solution. The ratio of dissociated molecules (n) to total molecules, *i.e.*, the sum of dissociated (n) plus nondissociated molecules (m) is the degree of dissociation (V):

$$V = \frac{n}{n + m}$$

A solution in which 30 per cent of the molecules are dissociated has a degree of dissociation equal to 0.3:

$$V = \frac{30}{30 + 70} = 0.3$$

The degree of dissociation increases with dilution and with rise in temperature; its greatest possible value is 1. The degree of dissociation should not be mistaken for the dissociation constant.

Dissociation constant. Mathematical analysis and experimental determination[1] have proved that the ratio of the product of the concentrations of the dissociated ions to the concentration of the nondissociated molecules is constant:

$$K = \frac{[A^-][C^+]}{[AC]}$$

in which K is the dissociation constant, $[A^-]$ the concentration of anions, $[C^+]$ the concentration of cations, and [AC] the concentration of nondissociated molecules. Each electrolyte has its own dissociation constant, which varies only with the temperature and remains unaltered at all concentrations.

The dissociation constant has particular significance in acids and alkalis, because their "strength," *i.e.*, acidity or alkalinity, depends on it.

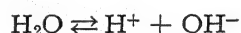
Acids dissociate into H^+ and the A^- corresponding to the particular acid. The dissociation constant of acids is expressed as follows:

$$K = \frac{[A^-][H^+]}{[AH]}$$

"Strong" acids (*e.g.*, HCl, H_2SO_4, HNO_3, etc.) have a high dissociation constant; therefore there are many H^+ in solution. "Weak" acids (*e.g.*, carbonic acid, acid phosphate, acetic acid, etc.) have a low dissociation constant and relatively few H^+ in solution.

Alkalis dissociate into hydroxyl ion (OH^-) and a cation; "strong" alkalis (*e.g.*, KOH, NaOH) have a high, and "weak" alkalis a low, dissociation constant.

Dissociation constant of water. Several experimental procedures have proved that pure water also dissociates into its constituent ions:

$$H_2O \rightleftarrows H^+ + OH^-$$

There are not many dissociated molecules; at 22°C. there are 18 gm. (1 mole) of water dissociated for every 10,000,000 liters. Therefore at 22°C. in every 10,000,000 liters of water there is 1 gram atom (*i.e.*, 1 gm.) H^+ and 1 gram atom (*i.e.*, 17 gm.) OH^-. In 1 liter of water there is 10^{-7} gram atom of H^+ and 10^{-7} gram atom of OH^-.

The dissociation constant of water is expressed as follows:

$$K = \frac{[H^+][OH^-]}{[H_2O]}$$

[1] Further details will be found in textbooks on physical chemistry or biophysics.

The concentration of nondissociated molecules is so high, compared to the concentration of dissociated molecules, and varies so little when the factors in the numerator change, that it can be considered constant without com-

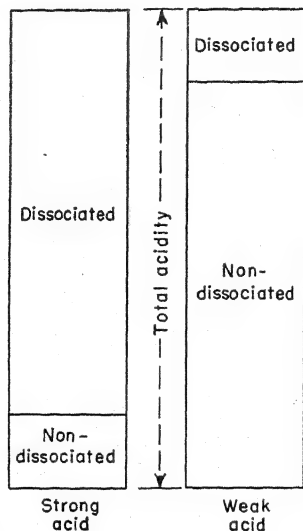



FIG. 142. Diagram illustrating total, potential, and real acidity. The weak acid has the same total acidity (titrated against a base) as the strong acid. The strong acid is highly dissociated and has a high real acidity. The weak acid is less highly dissociated and its real acidity is low. Potential acidity corresponds to the non-dissociated acid.

mitting any considerable error. Therefore

$$K = \frac{[H^+][OH^-]}{k}$$

where k is the concentration of nondissociated molecules. From this $K \times k = [H^+][OH^-]$, but the product of two constants is also a constant: $K \times k = Kw$, therefore the dissociation constant of water can be expressed as

$$Kw = [H^+][OH^-]$$

At 22°C. it is $10^{-7} \times 10^{-7}$ or 10^{-14}, *i.e.*, $Kw = 0.000\,000\,000\,000\,01$. The value of this constant varies with the temperature, since the degree of dissociation increases with the temperature, but it is usual when making calculations to refer to the dissociation constant at 22°C., *i.e.*, when $Kw = 10^{-14}$.

The dissociation constant of water has considerable significance. Neutral water is not water that does not contain H^+ and OH^-, but water in which there is the same concentration of both these ions. If NaCl is dissolved in water, most of it is dissociated into Na^+ and Cl^-, but the concentration of H^+ and OH^- does not vary, therefore the solution remains neutral.

If an acid or a base is added to neutral water a different process takes place. An acid is, by definition, a substance which releases H^+ when in solution. For example, if HCl is added to neutral water, the acid dissociates into H^+ and Cl^-, but the dissociation constant of the water remains unchanged. Let us suppose that HCl was added in sufficient quantity to give a H^+ concentration of 0.01 (*i.e.*, 10^{-2}); automatically the OH^- concentration shifts to 10^{-12} and $[H^+][OH^-] = 10^{-2} \times 10^{-12} = 10^{-14}$. If an alkali is added, *e.g.*, NaOH, in a quantity such that the concentration of OH^- is 0.1 (*i.e.*, 10^{-1}), the concentration of H^+ automatically shifts to 10^{-13}.

Therefore when the cH^+ is known the cOH^- is also known and acidity or alkalinity can be expressed in the concentration of either ion. It is usual to express it in the cH^+. Solutions in which $cH^+ = 10^{-7}$ are neutral. If this value increases, the solutions are acid; if it decreases, they are alkaline.

Real, potential, and total acidity. Equal volumes (*e.g.*, 10 cc.) of 0.1 *N* HCl and 0.1 *N* acetic acid have the same acidity when titrated with an alkali. Thus both require 10 cc. 0.1 *N* NaOH to be completely converted into NaCl and sodium acetate respectively. This acidity determined by volumetric titration is known as total acidity.

There is, however, a certain difference between the solutions of these acids; HCl is a "strong" acid, highly dissociated, and acetic acid is a "weak" acid, much less dissociated. A 0.1 *N* solution of HCl is 84 per cent dissociated; acetic acid in 0.1 *N* solution is only 1.36 per cent dissociated. The value of this real acidity, which varies for each acid, is given by the H^+ concentration. Figure 142 illustrates diagrammatically the acidity of strong and weak acids. Total acidity is the same for both acids, but real acidity is greater in the strong acids.

Potential acidity corresponds to the nondissociated fraction of the acid. When base is added the nondissociated molecules are progressively dissociated until the acid is completely neutralized.

The determination of real acidity (expressed

as cH^+) has great biological importance. For example, the intensity of the acid taste of a solution is due to its cH^+; sucrose in solution boiled in the presence of an acid is split (hydrolyzed) into fructose and glucose; the velocity of this reaction is proportional to the cH^+. Many other instances could be given in which the physiologic or toxic effects of an acid are due to the cH^+.

The pH nomenclature. To understand clearly the significance of the pH notation of cH^+ it is necessary to recall certain mathematical rules referring to logarithms.

The cH^+ of organic fluids and of alkaline solutions is extremely low; thus the cH^+ of water is 0.000 000 1, and that of 0.1 *N* NaOH is in the twelfth figure after the decimal point. The writing of these figures is cumbersome, and it is easy to make mistakes in reading them. Fortunately there are other, simpler ways of expressing them. The figures 100 and 1,000 can be written as 10^2 and 10^3 respectively. The exponents of 10 are the logarithms of the number represented; 2 is the logarithm of 100 and 3 the logarithm of 1,000. Two figures can be multiplied by adding their logarithms; the sum is the logarithm of the product. Thus

$$100 \times 1{,}000 = 10^2 \times 10^3 = 10^{2+3} = 10^5 = 100{,}000$$
$$\log 100 + \log 1{,}000 = \log 100{,}000$$

To divide one figure by another their logarithms are subtracted:

$$10{,}000 \div 100 = 100$$
$$10^4 \div 10^2 = 10^{4-2} = 10^2$$
$$\log 10{,}000 - \log 100 = \log 100$$

The log of 1 is 0, as is shown by the following calculation:

$$100 \div 100 = 1$$
$$100 \div 100 = 10^2 \div 10^2 = 10^{2-2} = 10^0$$
therefore $1 = 10^0$ and $\log 1 = 0$

The logarithms of figures in geometric progression are in arithmetic progression:

$\log 100 = 2$ or $100 = 10^2$
$\log 10 = 1$ or $10 = 10^1$
$\log 1 = 0$ or $1 = 10^0$
$\log 0.1 = -1$ or $0.1 = 10^{-1}$
$\log 0.01 = -2$ or $0.01 = 10^{-2}$
$\log 0.001 = -3$ or $0.001 = 10^{-3}$

Therefore quantities smaller than unity can be easily expressed as negative exponents of 10. This negative exponent will increase as the figures diminish. For example, the cH^+ of neutral water is 0.000 000 1, which can be more easily expressed as 10^{-7}. Not only multiples or submultiples of 10 can be represented in this way, but also other numbers. For example $2{,}300{,}000 = 2.3 \times 1{,}000{,}000 = 2.3 \times 10^6$. The logarithm of 2.3 is 0.36 (the two first decimal figures are sufficient for the accuracy of pH determinations). Therefore

$$2.3 \times 1{,}000{,}000 = 10^{0.36} \times 10^6 = 10^{0.36+6} = 10^{6.36}$$
and $\log 2{,}300{,}000 = 6.36$

Now let us take the case of a figure below unity, *e.g.*, a cH^+ of 0.000 002 3. This can be written in several other ways:

$$0.0000023 = 2.3 \times 0.000001 = 2.3 \times 10^{-6} = 10^{0.36} \times 10^{-6} = 10^{-5.64}$$
Since $0.36 - 6 = -5.64$
therefore $0.0000023 = 10^{-5.64}$

The exponent 5.64, without the negative sign, is the pH[1] corresponding to $cH^+ = 2.3 \times 10^{-6}$. As a general rule $cH^+ = 10^{-pH}$. In equations of this type the exponent of 10 is the log of the figure in the other term of the equation; therefore $-pH = \log cH^+$. Multiplying by -1 we have $pH = -\log cH^+$; also $pH = \log \frac{1}{cH^+}$ because the exponent of 10 changes sign on passing from the numerator to the denominator. Thus if $cH^+ = 0.001$, then $pH = \log \frac{1}{0.001}$, and as $\frac{1}{0.001} = 1{,}000$, $pH = \log 1{,}000 = 3$.

The pH, therefore, is the logarithm of the hydrogen ion concentration without the negative sign, and also the logarithm of the reciprocal of the cH^+.

The cH^+ of a solution is now always expressed by the pH notation introduced by Sörensen. The figure corresponding to the cH^+ is converted into an exponent of 10, and this exponent without the negative sign is the pH. Thus the cH^+ of water at 22°C. is $1 = 10^{-7}$, and the pH is 7.

This is the cH^+ concentration of water at 22°C.; therefore pH 7 is that of neutral water at this temperature. Dissociation increases with temperature, however, and at other temperatures the pH of neutral water is different; *e.g.*, at 18°C., $Kw = 0.64 \times 10^{-14}$. In neutral water, $[H^+] = [OH^-]$; therefore at 18°C. the concentration of both these ions is 0.8×10^{-7}, since $0.8 \times 10^{-7} \times 0.8 \times 10^{-7} = 0.64 \times 10^{-14}$.

[1] The symbol pH is formed by the initial of "power" (exponent) and the symbol for hydrogen.

The cH^+ of neutral water at 18°C. is therefore

$$0.8 \times 10^{-7} = 8 \times 10^{-8} = 10^{0.9} \times 10^{-8} = 10^{-7.1};$$

therefore the pH of neutral water at 18°C. is 7.1.

At body temperature (37°C.) the neutral pH is 6.78. When no mention is made of the temperature, the pH corresponds to that at 22°C., at which the neutral point is pH 7.

It is important to note that as real acidity increases the pH diminishes, because the value of the pH indicates the position of the significant figure after the decimal point. Thus for $cH^+ = 0.001$, $pH = 3$ and for $cH^+ = 0.01$, $pH = 2$, and as the significant figure is nearer to the decimal point the concentration increases. The pH indicates the dilution of H^+; a high pH means a high dilution, and vice versa.

Another result of this way of expressing cH^+ is that for every unit increase in the pH the H^+ are diluted 10 times.

The normal values of the pH of organic fluids are given in Table 26.

Table 26. Normal pH Values in Organic Fluids

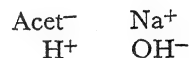
Fluid	*pH Value*
Gastric juice	1.0–3.0
Urine	4.8–7.4
Blood	7.3–7.45
Saliva	5.8–8.5
Bile	7.5–8.1
Pancreatic juice	7.3–8.1

Methods for determining pH. There are two methods most commonly used to determine the pH of a solution; one uses indicators, the other a potentiometer. The indicator method uses chemical substances that change color when the cH^+ of the solution changes. The pH at which the change of color takes place is characteristic for each indicator, and a series of these can be used for determining the different pH ranges.

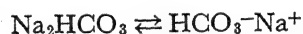
In the potentiometric method a platinum electrode (or a gold electrode plated with platinum) is saturated with hydrogen so that when placed in a fluid it behaves as a hydrogen electrode. When a zinc electrode is submerged in a solution containing zinc ions, a potential is established which is dependent on the zinc ion concentration of the solution; similarly with a hydrogen electrode a potential is established which is dependent on the cH^+ of the solution. This potential is measured with a potentiometer and the pH of the solution is then calculated.

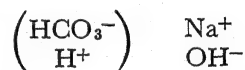
Buffer systems. Salts of a weak acid and a strong base (*e.g.*, sodium acetate) or of a strong acid and a weak base (*e.g.*, ammonium chloride) act as buffers. In the organism only buffers of the former type are found. Sodium acetate can be taken as an example. The $CH_3{\cdot}COO^-$ (acetate ion) will be expressed as $Acet^-$ and the nondissociated acetic acid as Acet H.

Buffer salts do not give a neutral solution as do salts of a strong acid and a strong base. For example, NaCl dissociates into Na^+ and Cl^-, but the H^+ and OH^- in the water of the solution do not vary. Buffer salts are also dissociated to a high degree when in solution, and as there are H^+ and OH^- in the solution, the situation can be represented as

$$\begin{matrix} Acet^- & Na^+ \\ H^+ & OH^- \end{matrix}$$

Reading the columns vertically it is seen there is acetic acid (a weak acid) and NaOH (a strong base). The weak acid is not considerably dissociated, and a high proportion of nondissociated acetic acid (Acet H) is found. The strong base on the other hand is considerably dissociated. Therefore in a solution of sodium acetate the OH^- predominates and the solution is alkaline. Of course, this alkalinity is not considerable and is not equivalent to a solution of NaOH, as might be thought by the example given. These physicochemical reactions are the result of equilibriums that can be calculated according to the mass law for each one of the electrolytes that takes part in them. These calculations will be found in treatises on physical chemistry or biophysics.

A very important buffer found in the organism is sodium bicarbonate, which is dissociated into Na^+ and bicarbonate ion:

$$Na_2HCO_3 \rightleftarrows HCO_3^-Na^+$$

As there are ions in the water of the solution, this can be represented by

$$\begin{pmatrix} HCO_3^- \\ H^+ \end{pmatrix} \quad \begin{matrix} Na^+ \\ OH^- \end{matrix}$$

The column on the left in parentheses represents the weak carbonic acid which is not much dissociated; and the column on the right the strong, highly dissociated base. An aqueous solution of $NaHCO_3$ is alkaline (turns litmus paper blue). This explains the paradoxical fact that an acid salt (sodium acid carbonate) gives an alkaline solution.

Many substances act as buffers in the organism; among the most important are sodium bicarbonate, disodium phosphate, proteins, etc.

A buffer system is obtained by adding the corresponding acid to the buffer, *i.e.*, it is a solution of a weak acid and its salt. An important property of buffers is made evident in buffer systems. For example, in a solution of acetic acid and sodium acetate, the acid dissociates according to its dissociation constant:

$$\frac{[Acet^-][H^+]}{[Acet\ H]} = K$$

whence $$[H^+] = K\frac{[Acet\ H]}{[Acet^-]}$$

Acetic acid is a weak acid; therefore it is not greatly dissociated. If sodium acetate, which has a considerable degree of dissociation, is added, the concentration of acetate ion ($Acet^-$) increases and a new dynamic equilibrium is established in which more of the $Acet^-$ is formed from sodium acetate and less from acetic acid. The dissociation of the acid diminishes and therefore the cH^+ also diminishes.

In this system the dissociation of $Acet^-$ is dependent on the degree of dissociation (δ) of the sodium acetate.

$$Acet^-] = \delta[Na\ Acet]$$

Therefore, substituting ($Acet^-$) in the former equation, we have

$$[H] = K\frac{[Acet\ H]}{\delta[Na\ Acet]}$$

The degree of dissociation varies with the concentration, but as there are no great changes in concentration in organic fluids, δ can be considered as constant. K/δ is therefore also a constant, known as K_1, whence

$$[H^+] = K_1\frac{[Acet\ H]}{[Na\ Acet]}$$

or, in general terms, $$[H^+] = K_1\frac{[acid]}{[salt]}$$

This is known as *Henderson's equation*, and it has great importance in physiology. It signifies that in a buffer system hydrogen ion concentration is dependent on the relative concentrations of the weak acid and its salt; or, conversely, that a change in the concentration of the acid or the salt modifies the hydrogen ion concentration.

Hasselbalch's equation expresses the same fact in terms of pH. Converting Henderson's equation into the reciprocal we have

$$\frac{1}{[H^+]} = \frac{1}{K_1} \times \frac{[salt]}{[acid]}$$

which can be expressed logarithmically as

$$\log\frac{1}{[H^+]} = \log\frac{1}{K_1} + \log\frac{[salt]}{[acid]}$$

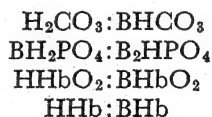
The first term of this equation is the pH; $\log\frac{1}{K_1}$ can be expressed as pK_1, so we have

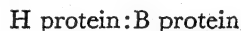
$$pH = pK_1 + \log\frac{[salt]}{[acid]}$$

This is Hasselbalch's equation; it has the same significance as Henderson's equation, from which it is obtained. Its importance will be seen when considering the acid-base equilibrium.

These systems are known as buffer systems because they modify the effects on the cH^+ of the addition of acid or base. For example, if 1 cc. 1 N HCl is added to 1 liter of water, a 0.001 N HCl solution is obtained. At this dilution the HCl is almost completely dissociated, so the cH^+ will be $0.001 = 10^{-3}$ and the pH = 3. If the same amount of HCl is added to a buffer system, *e.g.*, acetic acid–sodium acetate, the H^+ dissociated from the HCl are in the presence of an excess of $Acet^-$ dissociated from the sodium acetate, and a large number of nondissociated acetic acid molecules are formed. Only a few H^+ originated in the dissociation of acetic acid remain in solution, so there is little change in the real acidity of the solution.

There are many buffer systems in the organism contributing to the maintenance of a stable cH^+. The most important are the following (base, mainly Na and K, is expressed by the letter B):

$$H_2CO_3 : BHCO_3$$
$$BH_2PO_4 : B_2HPO_4$$
$$HHbO_2 : BHbO_2$$
$$HHb : BHb$$

Proteins also form a buffer system, since at the pH range of the organism they behave as weak acids:

H protein : B protein

When there are several buffer systems in a solution (as occurs in blood), changes in the concentration of one of the constituents of one system provoke changes in all the others, because of the shift in cH^+. Thus interdependent equilibriums between the different buffer systems are established.

BIBLIOGRAPHY

See references at the end of Chap. 30.

CHAPTER 30

The Transportation of Carbon Dioxide by the Blood

ROLE OF THE BLOOD PLASMA

If a strong acid is added to "true" plasma (see Chap. 31) and then it is placed in a vacuum, from 50 to 70 cc. of CO_2 is obtained from each 100 cc. of plasma.

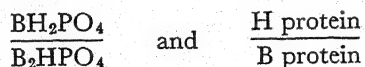
Approximately 5 per cent of the CO_2 is in solution in the form of carbonic acid (H_2CO_3); the rest forms an aqueous solution of bicarbonate. The former is known as free CO_2 and the latter as combined or fixed CO_2.

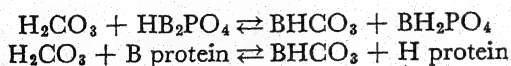
Applying Henderson's equation we have

$$[H^+] = K_1 \frac{[H_2CO_3]}{[BHCO_3]}$$

From this it can be seen that the cH^+ is dependent on the ratio of free CO_2 to combined CO_2, and conversely this ratio is dependent on the cH^+.

When arterial blood comes into contact with the tissues at the level of the capillaries, CO_2 enters the blood and the numerator in Henderson's equation tends to increase; therefore the cH^+, or real acidity of the blood, also tends to increase. Other buffer systems immediately come into action which check this tendency toward acidity, *e.g.*, the systems

$$\frac{BH_2PO_4}{B_2HPO_4} \quad \text{and} \quad \frac{\text{H protein}}{\text{B protein}}$$

The entrance of CO_2 and the shift toward acidity provoke the establishment of new equilibriums in these systems:

$$H_2CO_3 + HB_2PO_4 \rightleftarrows BHCO_3 + BH_2PO_4$$
$$H_2CO_3 + \text{B protein} \rightleftarrows BHCO_3 + \text{H protein}$$

An increase in CO_2 causes a shift toward the right in these reactions and therefore an increase in the bicarbonate of the plasma, which buffers the shift toward acidity in the acid-base equilibrium, *i.e.*, it controls the increase in cH^+. In the lung capillaries CO_2 passes from the blood to the alveolar air, and the changes in the plasma are opposite to those taking place in the tissue capillaries.

The buffer systems in the plasma are, however, of limited efficiency in the control of the acid-base equilibrium; whole blood is much more efficacious.

ROLE OF THE ERYTHROCYTES

Whole blood is collected and clotting prevented (*e.g.*, by adding sodium citrate); it is then brought to equilibrium with a gas mixture in which there is a high CO_2 tension. It is centrifuged under a layer of liquid paraffin to prevent gas exchanges at its surface. Plasma thus obtained has a high bicarbonate content. This fact was observed in 1867 by Zuntz. Under these conditions base has apparently passed from the cells to the plasma, but this is not so, because Hamburger has demonstrated that the erythrocyte membrane is not permeable to Na^+ and K^+. A rather complicated series of shifts take place, which is known as the Zuntz-Hamburger phenomenon.

The erythrocyte membrane is only slightly permeable to cations, such as Na^+, K^+, Ca^{++}, etc., but as Gürber showed in 1895, it is permeable to HCO_3^- and Cl^-. It is also permeable to H^+ and not permeable to hemoglobin.

The cells contain K^+ and hemoglobin, the latter being more acid when saturated with O_2

(HbO_2) than when in the reduced condition (Hb).[1] Therefore hemoglobin gives up cations (especially K^+) more readily than HbO_2.

Taking these facts into account a diagrammatic representation (Fig. 143) of what takes place can be made. This diagram only describes approximately the salient features and practical results of the changes in the different dynamic equilibriums. It is necessary to bear in mind that the liberation and fixation of ions are conditioned by interdependent equilibriums, which respond to the mass law.

As the blood passes through the tissue capillaries, its CO_2 concentration increases and HbO_2 diminishes, being converted into hemoglobin by loss of O_2. Hemoglobin is less acid and liberates K^+ more readily than HbO_2, fixing H^+ in its place. CO_2 entering the plasma combines with water and forms H_2CO_3, which dissociates (but not much) into HCO_3^- and H^+.[2] The H^+ enters the erythrocytes (and the cH^+ of the plasma diminishes) and is fixed by hemoglobin, taking the place of the K^+ set free. HCO_3^- also enters the red cells and combines with K^+, forming $KHCO_3$. In the plasma there is also dissociated NaCl, so Cl^- enters the red cells (Cl increases in the cells) and combines with K^+ into KCl; the Na^+ in plasma is balanced by the increase in HCO_3^-, and sodium bicarbonate increases. The increase in base observed in the plasma is therefore due not to a passage of base from the cells to the plasma, but to the passage of Cl^- from the plasma to the cells, which leaves an excess of cations in plasma. Therefore it would be more exact to say that there is a decrease in plasma anions (Cl^-) than that there is an increase in plasma cations (Na^+). Moreover, as CO_2 concentration increases in the blood, water passes from the plasma into the erythrocytes and the erythrocyte volume increases.

Changes in the opposite direction are observed in the lung capillaries. Oxygen enters the blood from the alveolar air and combines with hemoglobin, forming the more acid HbO_2, which sets free H^+; the H^+ diffuses out of the red cells into the plasma. Hemoglobin takes up K^+ from $KHCO_3$ and KCl. The anions of these salts, HCO_3^- and Cl^-, which had entered the cells as the blood changed from arterial to venous blood in its passage through the tissue capillaries, now return to the plasma. HCO_3^- is eliminated as CO_2 into the alveolar air, together with part of the CO_2 dissolved in the plasma. There is therefore an excess of Na^+, set free from the $NaHCO_3$; these Na^+ ions are balanced by the negative charges of the Cl^- ions diffusing out of the erythrocytes, which decrease in volume because water passes out of them into the plasma.

In tissues | In lung

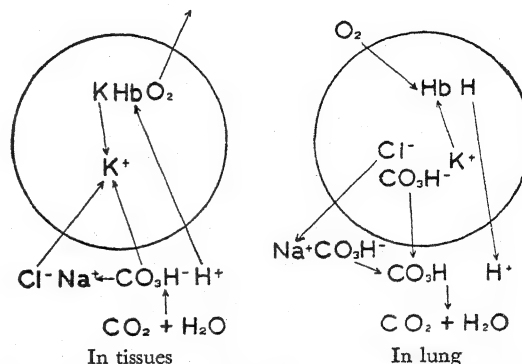
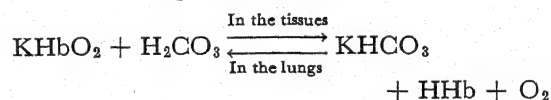



FIG. 143. Gaseous and ionic equilibrium between the erythrocytes and blood plasma (the Zuntz-Hamburger phenomenon). Explanation in text.

Hemoglobin plays an important, although indirect, part in the carriage of CO_2; reduced hemoglobin, which is less acid than HbO_2, favors the fixation of CO_2; HbO_2 formed in the lung capillaries favors the liberation of CO_2. This can be expressed as follows:

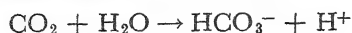
$$KHbO_2 + H_2CO_3 \underset{\text{In the lungs}}{\overset{\text{In the tissues}}{\rightleftharpoons}} KHCO_3 + HHb + O_2$$

Reduction of HbO_2 to Hb *ipso facto* allows the blood to fix a larger amount of CO_2. The importance of this factor in the carriage of CO_2 by the blood is great, since it is responsible for 50 per cent of the CO_2 carried from the tissues to the lungs. It must be understood that this does not mean half the total CO_2 in blood, but half the CO_2 that is to be eliminated by the lungs, *i.e.*, half the difference between the arterial and venous CO_2.

CO_2 is formed in the tissues by the oxidation of C atoms in metabolites and is eliminated as gaseous CO_2 by the lungs. A small amount

[1] The isoelectric point of hemoglobin is at pH 6.78, and that of HbO_2 at pH 6.6.

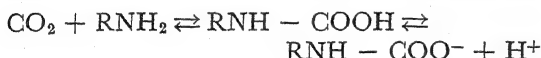
[2] The process is in truth more complicated because H_2CO_3 is a pseudoelectrolyte and exists only as dissolved CO_2 or dissociated H_2CO_3. For the sake of simplicity the process is described as if carbonic acid were not a pseudoelectrolyte.

circulates in simple solution but the greater part is in the state of bicarbonate ion. The following reaction takes place in the tissues:

$$CO_2 + H_2O \rightarrow HCO_3^- + H^+$$

The reverse occurs in the lungs.

At the pH of the blood, the formation of CO_2 from H_2CO_3 and the shift in the opposite direction take place at a very slow rate. The passage through the lung capillaries takes only approximately 1 sec.; therefore if H_2CO_3 in blood were in the same conditions as in an aqueous solution of the same pH, less than 1 per cent of the CO_2 eliminated could be set free in that short time. In the lung capillaries CO_2 is set free very rapidly, almost "explosively." This is due to two causes: (*a*) the formation of carbamino compounds with hemoglobin; (*b*) the existence of an enzyme, carbonic anhydrase, in the erythrocytes, which catalyzes the reaction $CO_2 + H_2O \rightleftarrows H_2CO_3$ in both directions.

Carbamino compounds. CO_2 (but not H_2CO_3, HCO_3^-, or CO_3^{--}) combines with amines and forms carbamino compounds. The carbamino reaction can be expressed as follows:

$$CO_2 + RNH_2 \rightleftarrows RNH - COOH \rightleftarrows RNH - COO^- + H^+$$

The velocity of this reaction in different conditions (pH, temperature, etc.) has been determined for many simple amines. Carbamino compounds do not precipitate with calcium and barium salts, so they can be easily separated from solutions that also contain carbonates or CO_2.

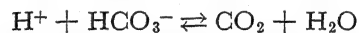
The following facts are evidence that a carbamino reaction takes place in the blood:

1. At a physiologic pH, solutions of pure hemoglobin fix CO_2 proportionally to the hemoglobin concentration; oxygenation of hemoglobin diminishes the amount of CO_2 taken up.
2. The CO_2Hb compound is not precipitated by $BaCl_2$, therefore CO_2 is neither in solution nor as bicarbonate ion.
3. The dissociation curve of CO_2Hb in function of time is the same as that of carbamino compounds of simple amines.
4. Curves of absorption and liberation of CO_2 by blood in which carbonic anhydrase has been inhibited by HCN (see further on) are similar to the curves observed with solutions of simple amines.
5. After precipitation of CO_2 and bicarbonate by $BaCl_2$ there is still a residual fraction of CO_2 proportional to the hemoglobin content.

Liberation of CO_2 from the carbamino compound of hemoglobin takes place at a rapid rate and is favored by oxygenation of hemoglobin.

Carbonic anhydrase. According to the classical theory, CO_2 is carried in the blood as free (dissolved) CO_2 and as bicarbonate. CO_2 in simple solution diffuses rapidly and can also be rapidly set free in the short time the blood takes to pass through the lung capillaries. On the other hand this theory gives no satisfactory explanation of the "explosive" liberation of CO_2 from bicarbonate that takes place in the lung, because at the pH of the blood CO_2 is set free from aqueous solutions of bicarbonate at a very slow rate. Therefore it was necessary to postulate that (*a*) CO_2 is at least in part carried in some other combination, not bicarbonate; (*b*) a catalyzer accelerates the liberation of CO_2 from bicarbonate.

Both these mechanisms have been demonstrated. The first is the formation of carbamino compounds explained in the preceding paragraph. The second is carried out by an enzyme in the erythrocytes, carbonic anhydrase, which accelerates the reaction

$$H^+ + HCO_3^- \rightleftarrows CO_2 + H_2O$$

A buffer system that contains H_2CO_3, at the pH of the blood, rapidly sets free CO_2 on addition of hemolyzed erythrocytes. The active fraction separated from the red cells has no hemoglobin, or hematin, and its activity is not due to catalase, peroxidase, or oxidase. The purified enzyme accelerates the liberation of CO_2 even when diluted 1 in 10,000,000. Carbonic anhydrase is inhibited by HCN, H_2S, CO, and several metal ions (Cu, Ag, Au, Zn, Hg).

The importance of the erythrocytes in the carriage of CO_2 is thus further enhanced, and the explanation of the Zuntz-Hamburger phenomenon must be completed as is illustrated in Fig. 144.

The total amount of CO_2 set free from the blood as it passes through the lungs[1] of a subject at rest can be divided approximately into the following fractions:

[1] Not the total amount of CO_2 in the blood; this has another significance, which will be considered in the next chapter when dealing with the alkali reserve.

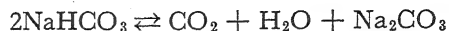
Dissolved CO_2 . . . . 10 per cent
Carbamino compounds of hemoglobin . . . . 20 per cent
Bicarbonate CO_2, set free by the action of carbonic anhydrase . . . . 70 per cent

CO_2 dissociation curves. A solution that contains CO_2, dissolved or combined (carbonates and bicarbonates), sets free CO_2 when a strong acid is added and it is then placed in a vacuum. The method most commonly used to determine the CO_2 content of organic fluids is that of van Slyke *et al.*, in which a measured volume of fluid is submitted to the effects of an acid and vacuum. The volume of all the gases set free is measured at atmospheric pressure;[1] a solution of NaOH is added to absorb the CO_2 and the gases are again measured. The difference between the two measurements gives the CO_2 content. The measurements should be corrected for NTP (*i.e.*, 0°C. and 760 mm. pressure).

CO_2 content can be expressed in several ways (*e.g.*, moles per liter), but in biology and medicine it is usual to give it in volumes of CO_2 per cent; *e.g.*, a blood CO_2 content of 55 volumes per cent means that 100 cc. of blood sets free 55 cc. CO_2 (at NTP) when submitted to the action of an acid and a vacuum.

[1] Van Slyke's manometric method is based on a different principle (see technical manuals).

CO_2 can be obtained from a solution by simple evacuation without adding acid, but in this case the carbonates dissolved do not set free CO_2. A solution of sodium bicarbonate sets free half its CO_2:

$$2NaHCO_3 \rightleftarrows CO_2 + H_2O + Na_2CO_3$$

This is a reversible reaction; therefore a solution of sodium carbonate in an atmosphere of CO_2 fixes this gas with the formation of $NaHCO_3$.

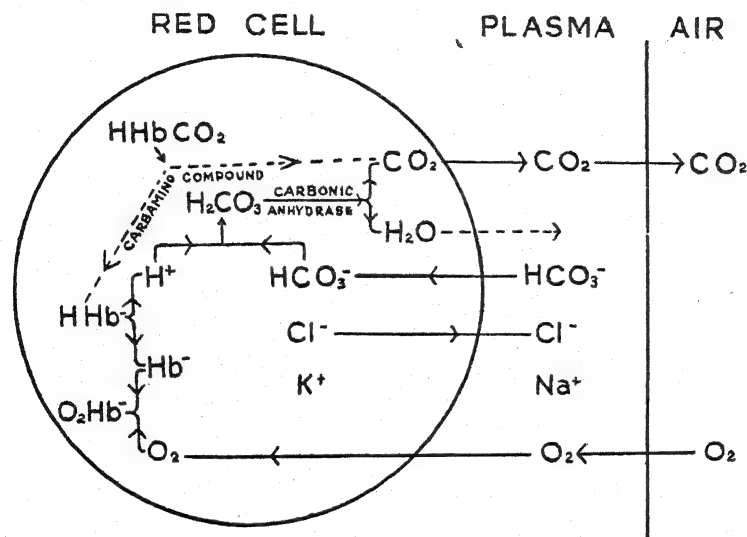



FIG. 144. Interchange of gases and ions between the erythrocytes, blood plasma, and alveolar air. Explanation in text. (*After Roughton.*)

Blood and other organic fluids in contact with gaseous mixtures of different CO_2 content take up or set free CO_2, according to the partial pressure of CO_2, until an equilibrium is established. For this reason the CO_2 content of blood has no precise significance if the CO_2 tension with which it is in equilibrium is not known.

A sample of blood (kept fluid by an anticoagulant) is divided into several fractions, each of which is brought to equilibrium with a gaseous mixture at different CO_2 partial pressures; then the CO_2 content of the blood of each fraction is determined. The results plotted on an *x-y* graph give the CO_2 dissociation curve. Examples of these curves are given in Fig. 145. The lower curve was obtained with blood brought into equilibrium with air of different CO_2 tensions; the upper curve with the same blood brought into equilibrium with mixtures of CO_2 and H_2. Hemoglobin in the blood corresponding to the lower curve was almost com-

pletely oxygenated (HbO_2), while in the blood of the upper curve it was almost completely reduced (Hb).

Several facts can be deduced from the examination of these curves. At the same CO_2 tension, blood with reduced hemoglobin (Hb)

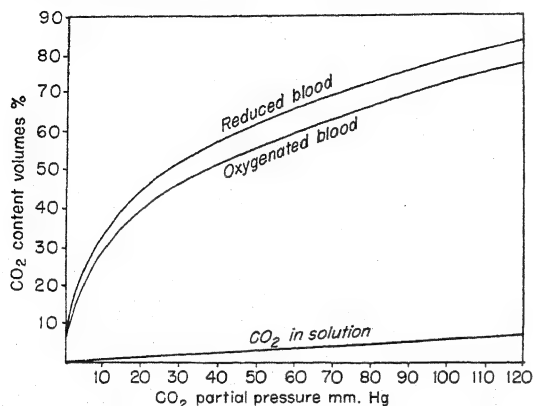
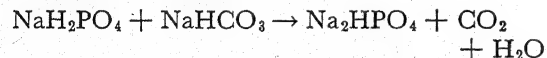



FIG. 145. CO_2 absorption curves of blood at different CO_2 partial pressures. The higher of the two curves corresponds to blood with reduced hemoglobin, and the lower to blood with oxygenated hemoglobin. The straight line at the bottom corresponds to CO_2 in solution.

fixes more CO_2 than blood with oxygenated hemoglobin (HbO_2). This is due to the lower acidity of hemoglobin, which permits it to fix a larger amount of H^+ and set free more K^+, and to the greater ease with which it enters into carbamino reactions.

Blood placed in a vacuum, or brought into equilibrium with a gas mixture without CO_2 (CO_2 partial pressure = 0), sets free all its CO_2, but plasma separated from the red cells does not. Therefore in whole blood—more precisely, in the erythrocytes—there must be a weak acid that can take care of the excess cations remaining after the elimination of CO_2 from bicarbonate. This weak acid that replaces HCO_3^- is hemoglobin.

Separated plasma also has several systems that can hold part of the excess Na^+, such as phosphate, proteinate, etc.; *e.g.*,

$$NaH_2PO_4 + NaHCO_3 \rightarrow Na_2HPO_4 + CO_2 + H_2O$$

These systems, however, are not sufficient to permit the liberation of all the CO_2.

The part of the CO_2 dissociation curve that is of greatest interest is that corresponding to the CO_2 tensions found in arterial and venous blood. As an example, let us take normal human blood with a CO_2 content of 52 volumes per cent in arterial and 57 volumes per cent in venous blood, and with 40 and 48 mm. Hg partial pressure respectively, and then examine Fig. 146, which is the section of Fig. 145 corresponding to the "physiologic" range of CO_2 tension.

The "arterial point" *A* on the lower curve corresponding to oxygenated blood lies at 40 mm. Hg CO_2 tension and 52 per cent CO_2 content. As blood passes through the capillaries, HbO_2 is reduced and CO_2 enters from the tissues, so that a "venous point" *V*, corresponding to 48 mm. Hg CO_2 tension and 57 per cent CO_2 content, is reached. On passing through the tissue capillaries, the blood has taken up 5 volumes per cent CO_2. If only the partial pressure of CO_2 varied, other circumstances remaining unaltered, the partial pressure would have had to rise to 52 mm. Hg (*B* on the lower curve) to obtain the increase of 5 volumes per cent in the CO_2 content. If this had taken place, dissolved CO_2 would have increased, and according to Henderson's equation, the cH^+ would

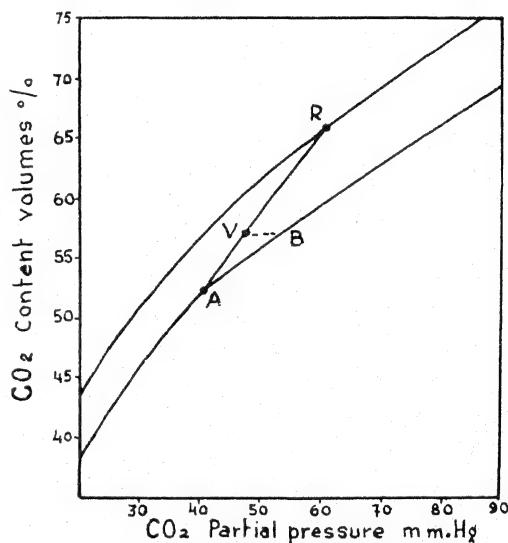



FIG. 146. Normal range of the curve in Fig. 145. Description in text.

have also risen. Partial reduction of hemoglobin causes the venous point to be at a lower CO_2 tension, therefore hemoglobin plays a part in maintaining the constancy of the acid-base equilibrium in the blood, damping the oscillations in pH.

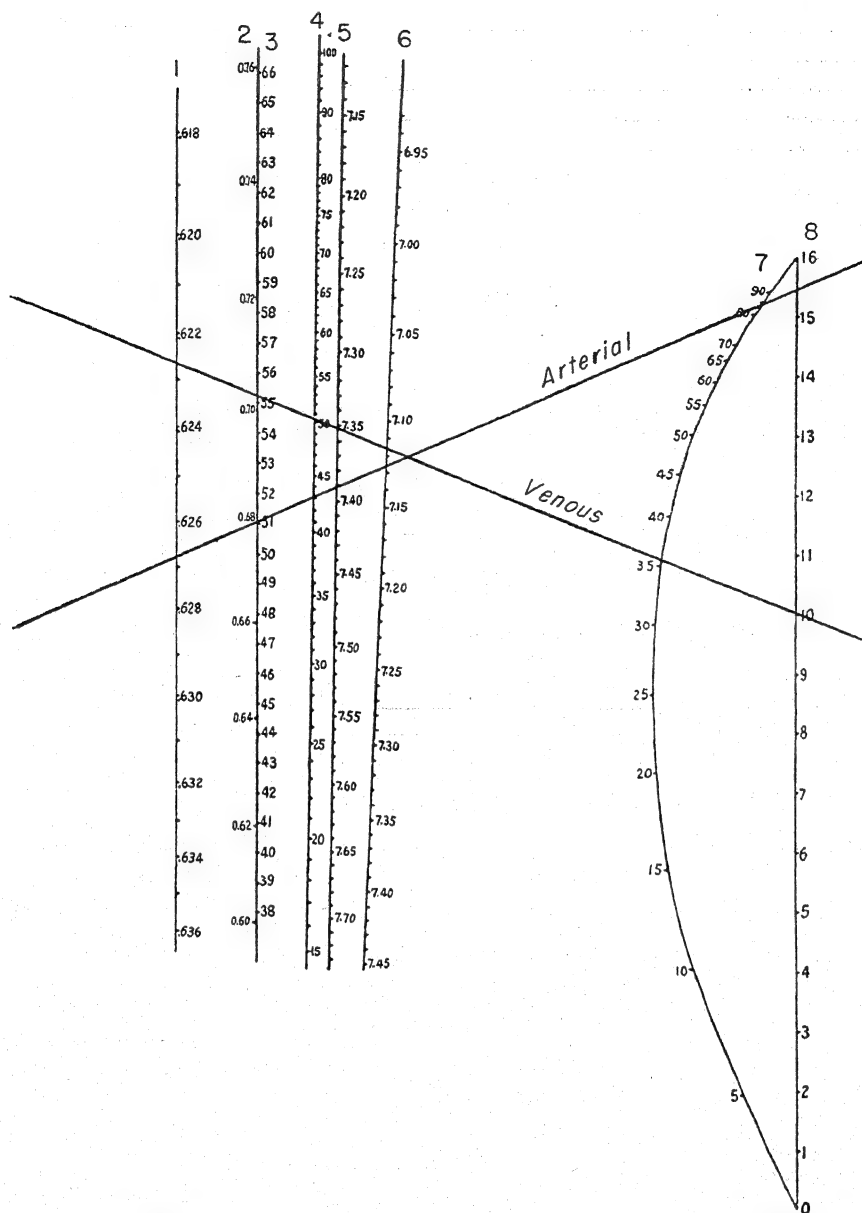



FIG. 147. Henderson's nomogram of human blood. Eight interdependent variables are represented in order from left to right: 1, ratio of plasma volume to total volume; 2, ratio of cell Cl to plasma Cl; 3, total CO_2 in volumes per cent; 4, CO_2 pressure in mm. Hg; 5, pH of plasma; 6, pH of erythrocytes; 7, O_2 pressure in mm. Hg; 8, HbO_2 expressed as volumes of O_2 per cent. (*Henderson, L. J., "Blood," Yale University Press, New Haven, 1928.*)

INTERDEPENDENT EQUILIBRIUMS. HENDERSON'S NOMOGRAMS

The student can now understand how changes in one component of the blood cause changes in many closely interdependent equilibriums. L. J. Henderson has made a particular study of these equilibriums. He considers that between both phases of the blood (plasma and cells) there are eight components, four of which can be distributed in varying proportions between both phases (water, carbonic acid, chloride, and oxygen); two are found only in plasma (plasma proteins and plasma base); and two are found only in the cells (hemoglobin and cell base).

Two of these components, oxygen and carbonic

acid, are independent variables; changes in the other components are subject to variations in these two. Changes in the eight components cause changes in other variables such as pH (dependent on the variables free CO_2 and combined CO_2), the Donnan equilibrium, the red cell volume, etc.

Henderson has determined experimentally the interdependence of these variables and expressed their relationship mathematically. Nomograms (Greek, *νομος*, law, and *γράμμα*, written) can be constructed so that the different variables are represented in a system of coordinates (D'Ocagne's nomograms) and their relationship rapidly understood. The number of variables in some of Henderson's nomograms is 18; in the one reproduced in Fig. 147, eight interdependent variables are represented.

THE ROLE OF CO_2 IN THE BODY

CO_2 is the end product of carbon metabolism in the animal organism, and it is eliminated as a waste product. Nevertheless some remains so as to maintain a certain level of CO_2 concentration in the fluids and tissues, which is of great physiologic importance. CO_2 combined with Na in $NaHCO_3$ forms the alkali reserve; furthermore a certain amount of free CO_2 is necessary for an adequate cH^+. Therefore excessive elimination of CO_2 may cause severe disturbance, *e.g.*, alkalosis (see "The Alkali Reserve," Chap. 31, and Chap. 32, Regulation of Respiration.)

CO_2 also plays an important part in organic synthesis. In plants it is the only source of C. The activity of chlorophyll and energy obtained from sunlight are the means used by plants to reduce CO_2 and thus have C for the synthesis of the organic compounds that make up the plant (cellulose, starch, proteins, oils, alkaloids, etc.).

Using a radioactive carbon isotope (C^{11}) to label CO_2, it has been possible to demonstrate that CO_2 gives up C in the synthesis of many compounds by bacteria or animal tissues, *e.g.*, methanol; urea; propanol; formic, acetic, pyruvic, lactic, propionic, oxalacetic, succinic, fumaric, malic, butyric, α-ketoglutaric, oxalsuccinic, and citric acids; and glycogen (Wood). In rats injected with sodium bicarbonate prepared with C^{11}, 13 per cent of the carbon in liver glycogen was found to be C^{11}.

BIBLIOGRAPHY

EVANS, E. A., JR., Carbon Dioxide Utilization in Animal Tissues, *Science*, **96**, 25, 1942.

HENDERSON, L. J., "Blood. A Study in General Physiology," Yale University Press, New Haven, 1928.

LOVATT-EVANS, C., in W. H. Newton, "Recent Advances in Physiology," 4th ed. and 6th ed., J. & A. Churchill, London, 1930 and 1939.

MENDIVE, J. R., Anhydrasa Carbónica, *Medicina, Buenos Aires*, 3, 366, 1943.

ROUGHTON, F. J. W., Recent Work on Carbon Dioxide Transport by the Blood, *Physiol. Rev.*, **15**, 241, 1935.

VENNESLAND, B., A. K. SOLOMON, J. M. BUCHANAN, and A. B. HASTINGS, Glycogen Formation from Glucose in the Presence of Radioactive Carbon Dioxide, *J. Biol. Chem.*, **142**, 379, 1942.

WOOD, H. G., The Fixation of Carbon Dioxide and the Interrelationship of the Tricarboxylic Acid Cycle. *Physiol. Rev.*, **26**, 198, 1946.

CHAPTER 31

The Acid-Base Equilibrium of the Organism

THE ALKALI RESERVE

In an aqueous solution of electrolytes the total positive charges are equal to the total negative charges; in other words, the total concentration of positive equivalents is equal to the total concentration of negative equivalents. Thus the electric neutrality of the solution is maintained. For example, in a NaCl solution, for each Na^+ there is one Cl^-; in a $CaCl_2$ solution, for each Ca^{++} there are two Cl^-. One equivalent is equal to one mole for monovalent ions; but for polyvalent ions one equivalent is equal to one mole divided by the number of valences. Thus one mole of Ca^{++} weighs 40 gm., and one equivalent of Ca^{++} weighs $40/2 = 20$. Therefore 20 gm. Ca^{++} is the equivalent of 35.5 gm. Cl^-, *i.e.*, in a solution containing these amounts of Ca and Cl there is the same number of positive charges (Ca^{++}) as of negative charges (Cl^-).

Electric neutrality, or electric equilibrium, is also maintained in complex solutions of electrolytes, such as blood plasma. The part taken by each of the ions is established by determining their concentration by chemical analysis, and their equivalents (usually milliequivalents when working with blood plasma) are obtained by applying the following formula:

$$\text{mEq. per liter} = \frac{\text{Mg. per liter} \times \text{valence}}{\text{Molecular weight}}$$

The concentration of the principal ions in plasma is given in Table 27 (Fig. 148). In this table and figure it can be seen that the negative charges are carried by strong acids (HCl, H_2SO_4) and by weak acids (H_2CO_3, acid phosphate and proteinates). The former, because of their strength, cannot be displaced from their base by acids formed in the course of the normal or abnormal metabolism of the body; the latter, on the other hand, can be displaced. For example, during strenuous muscular exercise the concentration of lactic acid in the blood increases considerably; this acid takes Na from $NaHCO_3$, and thus sets free H_2CO_3, which is eliminated as CO_2 by the lungs. There is, therefore, an alkali reserve in the blood that can neutralize the acid valences formed in the body. This alkali reserve is made up by the base of weak acid salts. The most important of these acids is carbonic acid, owing to the ease with which it is eliminated as CO_2 by the lungs. The alkali reserve of the blood is usually measured by determining the bicarbonate content of the plasma.

Table 27. Normal Content of Positive (Base) and Negative (Acid) Ions in Human Plasma

Base ions	*Content, mEq. per liter*		Acid ions	*Content, mEq. per liter*	
	Range	*Average*		*Range*	*Average*
Na^+	136.0–144.0	142	HCO_3^-	25.0–27.0	27
K^+	4.0–5.0	5	Cl^-	98.0–106.0	103
Ca^{++}	4.8–5.2	5	HPO_4^{--}	1.8–2.5	2
Mg^{++}	2.5–3.5	3	SO_4^{--}	0.8–1.2	1
			Organic	4.0–8.0	6
			Proteins	14.0–18.0	16
Total		155	Total		155

The bicarbonate content of plasma can be expressed in several ways, *e.g.*, in milligrams per cent or in millimoles per liter, but it is customary

to express it in centimeters (or volumes) of CO_2 per cent, because it is determined by the volume of CO_2 set free from the plasma by a strong acid *in vacuo*. The normal value lies between 50 and 70 volumes CO_2 per cent.

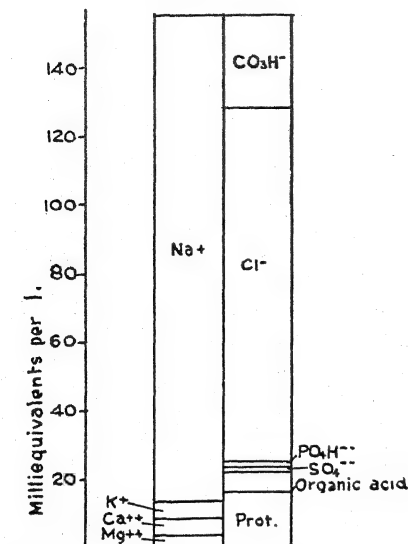



Fig. 148. Diagram of average content of anions and cations in human blood expressed in milliequivalents per liter.

Determination of the alkali reserve. The ease with which CO_2 diffuses out of the plasma into the air should be kept in mind when measuring the alkali reserve, because if adequate precautions are not taken the balance between the CO_2 in the plasma and that in the cells is disturbed, and errors are thus committed. The significance of certain terms in common use should be clearly understood.

True plasma is obtained from blood under conditions in which no changes in its gaseous content can occur. The blood must be centrifuged under a layer of paraffin oil so that gas cannot escape from the surface into the air, thus causing an interchange of ions between erythrocytes and plasma. After centrifuging, the plasma is pipeted off and placed in a saturating vessel. This true plasma, obtained from venous or arterial blood, can be equilibrated with the subject's own alveolar air; it is then called "regularized plasma." If it is equilibrated with a gaseous mixture containing 5.5 per cent CO_2, corresponding to a CO_2 tension of 40 mm. Hg, which is the average CO_2 tension in alveolar air of man, it is called "reduced plasma."

The following procedure is most commonly used: About 12 cc. of blood is drawn from a vein in the arm, taking care to avoid stasis and the accumulation of CO_2 in the blood, as this would cause an increase in plasma bicarbonate due to the chloride shift between the red cells and plasma, as explained in Chap. 30. The blood is collected in a syringe containing a small number of sodium oxalate or citrate crystals to prevent clotting, and is passed into a centrifuge tube under a layer of liquid paraffin. After centrifuging, 4 or 5 cc. is pipeted off, placed in a saturation flask (Fig. 149), and equilibrated with CO_2 at the same tension as in arterial blood by filling the flask with the subject's own alveolar air or a gas mixture with 5.5 per cent CO_2. In the first case the subject makes a forced expiration (after a normal inspiration) into the saturation flask through another flask containing glass beads to retain the moisture in the expired air. The saturation flask is then closed at both ends. If 5.5 per cent CO_2 is used, about 1 liter of the gas mixture is passed through the flask. The flask is gently

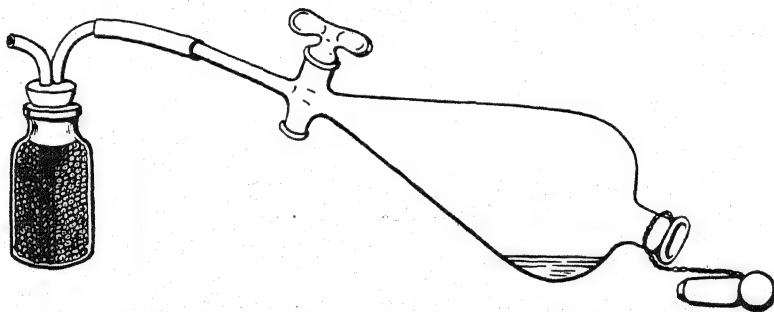



Fig. 149. Saturation flask. A few cubic centimeters of blood is placed in the flask and this is then filled with a gas mixture of known composition with which the blood is to be brought into gaseous equilibrium. The flask is closed at both ends and is gently rotated so as to distribute the blood along its walls, thus increasing the surface of contact between gas and blood. The bottle with glass beads is used to retain water vapor when the flask is filled with expired air; the bottle is disconnected after closing the flask by turning the stopcock.

rotated during 2 min. so that there is a large surface of contact between the air and plasma. Afterward 3 cc. of plasma is placed in a van Slyke gas apparatus, where CO_2 and other gases are set free from the plasma by adding a strong acid and producing a vacuum. The gas is brought to atmospheric pressure and measured, making the corrections for NTP. Not only CO_2 but also N_2 and O_2 are thus liberated, but the two latter are fairly constant, so the part corresponding to CO_2 can be read off in especially constructed tables. Technical manuals should be consulted for further details and for van Slyke and Neill's manometric method.

The CO_2 volume thus measured corresponds to the bicarbonate in plasma and is also called the carbon dioxide "combining power," or "capacity," of the plasma. This volume calculated for 100 cc. of plasma is the usual way of expressing the alkali reserve of the plasma. The normal figure is 50 to 70 volumes per cent. Unless special mention is made of the use of whole blood, lymph, pleural effusion, etc., it is understood that this figure corresponds to a determination made on venous blood plasma.

The alkali reserve is one of the constant properties of the organism, in the same way as the body temperature, blood-sugar level, etc. In normal conditions it does not vary to any considerable extent.

THE ACID-BASE EQUILIBRIUM

Two constants have importance in the study of the acid-base equilibrium, the hydrogen ion concentration, expressed by the pH, and the alkali reserve. These can vary independently of each other, since the pH is not dependent on the bicarbonate content but on the ratio of bicarbonate to carbonic acid.

In a buffer system consisting of bicarbonate and carbonic acid, the pH is conditioned as follows according to Hasselbalch's equation:

$$\mathrm{pH} = pK_1 + \log \frac{[\text{Bicarbonate}]}{[\text{Carbonic acid}]}$$

As bicarbonate is measured by the CO_2 set free, the last part of the second member of the equation can be replaced by the logarithm of

$$\frac{[\text{Combined } CO_2]}{[CO_2 \text{ in solution}]}$$

Since $pK_1 = 6.10$, if two of the remaining variables are known the third can be calculated.

The amount of CO_2 in solution is dependent on its partial pressure, which can be measured in the alveolar air. The absorption coefficient of CO_2 in plasma at 38°C. is 0.51. If the CO_2 partial pressure is 40 mm. Hg, the volume of CO_2 dissolved in 100 cc. at a barometric pressure of 760 mm. will be

$$40/760 \times 100 \times 0.51 = 2.68 \text{ volumes per cent}$$

If the total CO_2 in plasma amounts to 63 volumes per cent, and the CO_2 in solution is 3 volumes per cent, there will be 60 volumes per cent combined CO_2; the ratio in Hasselbalch's equation will be $60/3 = 20$. But $\log 20 = 1.30$; therefore $\mathrm{pH} = 6.1 + 1.30 = 7.4$. In normal conditions for each volume of CO_2 in solution there are approximately 20 volumes of combined CO_2, *i.e.*, 5 volumes of free CO_2 for every 100 volumes of combined CO_2. The volumes of CO_2 in solution per 100 volumes of combined CO_2 at different pH levels are given in Table 28.

Table 28. Free CO_2 for Each 100 Volumes of Combined CO_2 at Different pH

pH	*Free CO_2, Volumes*
7.0	12.5
7.3	6.3
7.4	5
7.5	4
7.8	2

The pH of the blood can therefore be determined by this indirect method, but it is not usually adopted because it is more complicated than the potentiometric and indicator methods.

A decrease in plasma bicarbonate is not accompanied by a change in pH if enough CO_2 is eliminated by the lungs to maintain the ratio of 1:20 between free and combined CO_2. This condition is known as compensated acidosis (see page 285). In a similar way ingestion of bicarbonate in sufficient quantities to increase its concentration in the blood does not modify the pH if at the same time enough CO_2 is retained to maintain a 1:20 ratio between free and combined CO_2. In this case there is compensated alkalosis.

These facts show the importance of pulmonary ventilation and the excitability of the respiratory center in the maintenance of a normal acid-base equilibrium and in some conditions in which it is disturbed.

The acid that is formed in greatest amounts in the body as a result of metabolic activity is

carbonic acid; from 400 to 460 liters, or 800 to 900 gm. of CO_2, is produced daily. These large amounts, however, do not disturb the normal acid-base balance because of the ease with which the lungs eliminate CO_2 as it is formed.

librium. The alkali reserve and the pH can be normal, or high, or low, and by the combination of these conditions of each of two factors, nine different cases are possible. They will be discussed in the following pages. The existence of

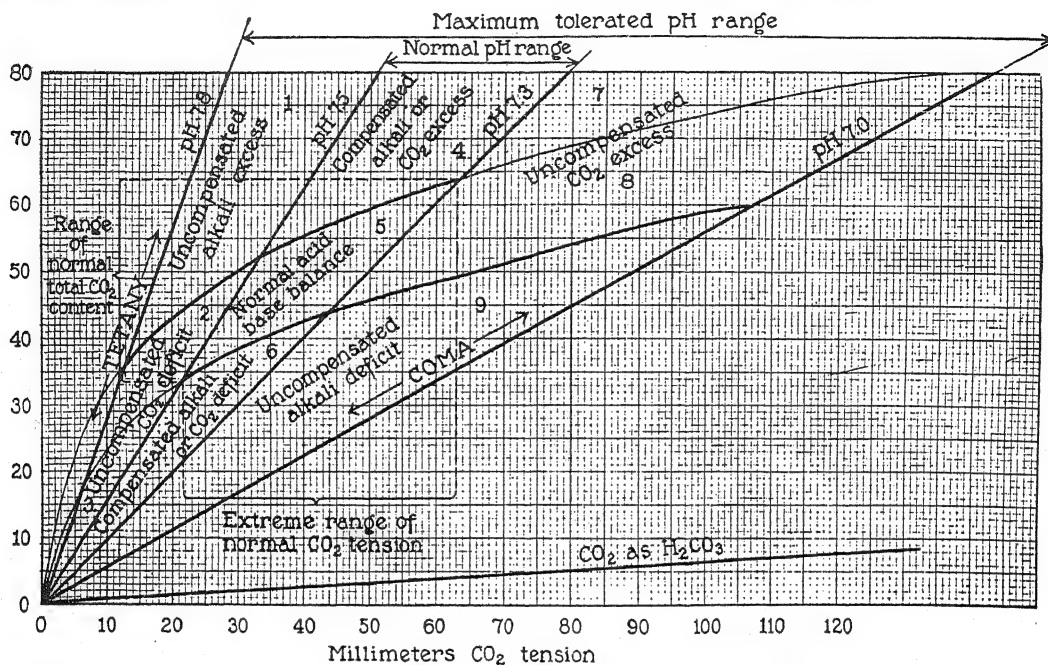



FIG. 150. Normal and pathologic variations in $BHCO_3$, H_2CO_3, CO_2 pressure, and pH in total human blood obtained from a resting subject at sea level. The curves represent CO_2 dissociation; the upper curve corresponds to reduced blood and the lower one to oxygenated blood. The straight line in the lower part of the diagram represents the CO_2 in solution, which is directly proportional to the partial pressure of CO_2. The other straight lines represent the pH calculated from the ratio of $BHCO_3$ to H_2CO_3 (combined CO_2 to free CO_2) of Hasselbalch's equation. The total combined CO_2 corresponding to any point on the pH lines is obtained by subtracting CO_2 in solution from total CO_2. For example, on the line corresponding to pH 7.0 at 107 mm. Hg CO_2 partial pressure, the total combined CO_2 is $60 - 6.7 = 53.3$. The ratio of dissolved CO_2 to combined CO_2 is 12.5 per cent. This ratio corresponds to pH 7.0 according to Hasselbalch's equation. This ratio is constant for all the points on the line; therefore the pH lines are straight ones. The nine areas separated by the different lines and curves are described in the text. (*Van Slyke, D. D., J. Biol. Chem., vol. 48, p. 157, 1921.*)

Venous blood is slightly more acid than arterial blood; there is a difference in pH of 0.02 between the two. This slight difference would be much larger if it were not for the mechanism described as the Zuntz-Hamburger phenomenon (Chap. 30).

The normal range of pH is from 7.3 to 7.5; extreme variations compatible with life are 7.0 and 7.8.

Disturbances in the acid-base equilibrium. Acidosis and alkalosis. The alkali reserve and the pH—two factors that vary independently of each other—should be determined when considering disturbances in the acid-base equilibrium. The alkali reserve and the pH can be normal, or high, or low, and by the combination of these conditions of each of two factors, nine different cases are possible. They will be discussed in the following pages. The existence of these two variables has caused some confusion as to the exact significance of "acidosis" and "alkalosis." The term "acidosis" does not mean that the blood is acid, since a pH below 7 is incompatible with life. Used in a less precise way, it signifies a decrease in pH or the alkali reserve, and "alkalosis" signifies the opposite conditions. This ambiguous use of the word "acidosis" occurs in van Slyke's definition; he considers that acidosis is due to the formation or absorption of acids at a rate that exceeds their rate of elimination, so that either the hydrogen ion concentration in the blood increases, or the alkali reserve decreases below the normal range.

To avoid confusion some authors do not employ the terms "acidosis" and "alkalosis," but as they are in common use, it is necessary to give them a definite meaning. They may therefore be defined as variations of the alkali reserve outside the normal range. If in acidosis, thus defined, the pH is normal (owing to an adequate decrease in free CO_2) the condition is one of compensated acidosis; if the pH falls below 7.3, there is uncompensated acidosis. The corresponding opposite states are known as compensated and uncompensated alkalosis.

The nine possible states produced by the three variants in the two factors have been described by van Slyke, who calls them "alkali or CO_2 excess, compensated or uncompensated, with high or low pH" (Figs. 150 and 151).

Possible variations in acid-base equilibrium (according to van Slyke).[1] *Area* 1. *Uncompensated alkali excess.* In this condition, bicarbonate ($NaHCO_3$) is increased without a corresponding increase in H_2CO_3, so the pH rises above the normal. This condition can be produced by excessive ingestion of sodium bicarbonate, or by the loss of HCl as a result of repeated vomiting or stomach washing. If the pH rises sufficiently, tetany may appear. Compensation is brought about by a decrease in active pulmonary ventilation, which causes retention of CO_2, and an increase in the renal excretion of bicarbonate and alkaline phosphate.

Areas 2 *and* 3. *Uncompensated CO_2 deficit.* Free $CO_2(H_2CO_3)$ is decreased without a proportionate decrease in combined $CO_2(NaHCO_3)$, so the pH increases. This disturbance is provoked by hyperpnea, such as is caused by lack of oxygen at high altitudes, or by hot baths, or by fever. If the pH rises sufficiently, tetany appears. Compensation takes place by the accumulation of acid metabolites, and by a decrease in the acidity or the alkalinization of the urine. Sometimes this condition is called "respiratory alkalosis" because it is due primarily to loss of free CO_2, which is eliminated by the lungs as a gas.

Area 4. *Compensated CO_2 or alkali excess.* In this condition there is a parallel rise in free carbon dioxide (H_2CO_3) and combined carbon dioxide ($NaHCO_3$); there is a high alkali reserve and the pH is within the normal range (*i.e.*, the disturbance is compensated). It is produced by the same causes that provoke the disturbance in area 1, but in a lesser degree, *i.e.*, ingestion of sodium bicarbonate with subsequent CO_2 retention, or primary retention of CO_2 as occurs in certain chronic diseases of the lung that hinder the gaseous exchange (emphysema) with secondary retention of $NaHCO_3$. If the disturbance is due to pulmonary disease, there is cyanosis due to insufficient oxygen absorption.

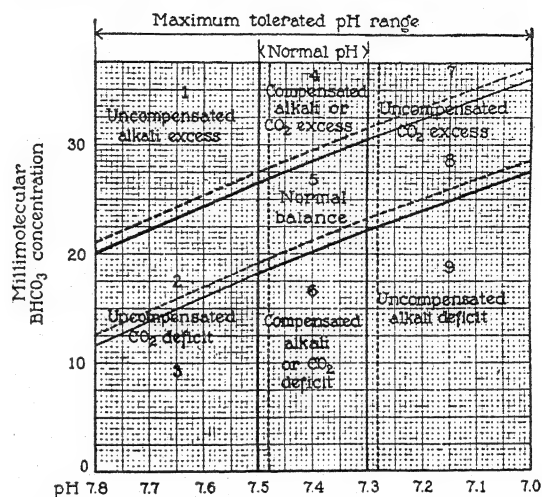



FIG. 151. Normal and pathologic variations in bicarbonate and pH of serum or oxalated plasma. This figure has the same significance as Fig. 150, but the pH is represented so as to facilitate distinction of the nine areas described in the text. Solid lines correspond to arterial blood and broken lines to venous blood. CO_2 is expressed not in volumes per cent but in millimolar concentration. As 1 mole of any gas has a volume of 22.4 liters (NTP), 1 millimole has a volume of 22.4 cc.; therefore a millimolar concentration of 1 corresponds to 22.4 cc. per liter, or 2.24 cc. per cent. (*Van Slyke, D. D., J. Biol. Chem., vol. 48, p. 157, 1921.*)

Area 5. *Normal acid-base equilibrium.* Combined (alkali reserve) and free CO_2 and the pH are normal at sea levels; at high altitudes the conditions vary (Chap. 33).

Area 6. *Compensated alkali or CO_2 deficit.* There is a proportionate decrease in combined and free CO_2; therefore the pH is normal. This condition can be initiated either by a deficiency of alkali or by a deficiency of CO_2. Alkali deficit can be due to excessive production of fixed acid metabolites (*e.g.*, ketone bodies), to retarded elimination of fixed acids (nephritis), to intoxication by acids that are not metabolized (HCl, H_2SO_4), to acidogenous salts,[1] or to excessive excretion of base, such as

[1] VAN SLYKE, D. D., Studies of Acidosis. The Normal and Abnormal Variations in the Acid-base Balance of the Blood, *J. Biol. Chem.*, **48,** 153, 1921.

[1] The more commonly used acidogenous salts are NH_4Cl and $CaCl_2$. The NH_4 is converted into urea, which is eliminated; thus acid Cl^- is released. In the

occurs in certain infantile diarrheas and in biliary and pancreatic fistulas.

An initial CO_2 deficit can be observed at high altitudes. It is due to hyperpnea (the typical sign of this condition) caused by the decrease in oxygen partial pressure.

When the initial cause is alkali deficit, it is compensated by an increase in the elimination of acids and ammonia in the urine (unless renal iniury prevents this) and a decrease in CO_2 partial pressure in alveolar air. When the initial cause is CO_2 deficit, there is an excess elimination of base through the kidneys until the free and combined CO_2 are in normal proportions. In high altitudes this is the compensatory reaction, and persons acclimatized to high altitudes have an alkali reserve below the normal one for sea level.

Areas 7 *and* 8. *Uncompensated CO_2 excess.* There is an increase in free CO_2 without a corresponding increase in base; therefore the pH is low. This condition is also known as "respiratory acidosis," because the initial disturbance is retention of free CO_2. It can be observed in pneumonia (a disease in which the respiratory surface of the lung is considerably diminished) in cases of intoxication by depressants of the respiratory center (*e.g.*, morphine), and in uncompensated cardiac insufficiency. Dyspnea is a prominent symptom of this state. Compensatory mechanisms are increased respiratory rate, renal excretion of acid metabolites with increased formation of ammonia, and perhaps diffusion of acids from the blood into the tissues.

Area 9. *Uncompensated alkali deficit.* There is a considerable decrease in blood bicarbonate, without a proportionate decrease in free CO_2; therefore the pH is low. This condition is observed in advanced diabetic coma, in the final stages of renal acidosis, in very deep anesthesia, and in certain forms of cardiac insufficiency. There is dyspnea, and compensatory reactions are the same as in uncompensated CO_2 excess, unless the condition is of renal origin, in which case the kidney does not take part in compensation.

Respiratory alkalosis and acidosis. Disturbances in pH due to variations in free CO_2 can cause acidosis (retention of CO_2) or alkalosis (excess elimination of CO_2). In the first case the condition corresponds to areas 7 and 8 of van Slyke's diagram, and in the second to areas 2 and 3. In these states the deficit or excess is due primarily to variations in free (gaseous) CO_2, represented in Hasselbalch's equation. For this reason these types of disturbance are known as respiratory acidosis or alkalosis. If acidosis is due initially to uncompensated loss of alkali (*i.e.*, a decrease in fixed CO_2), as in conditions corresponding to area 9, it is sometimes known as "metabolic acidosis."

case of $CaCl_2$ the cation is eliminated in large part as $CaCO_3$ through the intestine, and Cl^- is set free.

Changes in acid-base equilibrium are the results, or signs, of more or less serious conditions; they are not always a fundamental cause of the seriousness of the patient's condition. Two examples will make this clear. In diabetic coma there may be a considerable decrease in the alkali reserve; this can be increased by ingestion of sodium bicarbonate, thus reestablishing a normal acid-base equilibrium, but the seriousness of the patient's general condition remains unchanged. In deep asphyxia, in spite of the "acidosis" due to CO_2 retention, the appropriate treatment is to make the patient breathe a mixture of O_2 and CO_2 with the object of stimulating the respiratory center, thus alleviating the general condition.

MECHANISMS FOR THE REGULATION OF THE ACID-BASE EQUILIBRIUM

The normal range of blood pH is 7.3 to 7.45, and that of the alkali reserve is 50 to 70 volumes CO_2 per cent. Within these normal limits there are daily variations, the most important of which are due to muscular exercise (increased production of CO_2 and lactic acid) and to digestive processes (increase of the alkali reserve due to the secretion of HCl into the gastric juice). Besides these changes, the body is continuously producing acid or alkaline substances in the course of, and as a result of, the metabolism of foodstuffs.

Acid substances are produced in the following ways:

1. The oxidation of carbon gives rise to CO_2.
2. Proteins break down to amino acids, which are either resynthesized into protein or oxidized. Amino acids with sulfur in their molecule give sulfuric acid as a final product of disintegration.
3. Phospholipids and nucleoproteins give phosphoric acid; nucleoproteins also give rise to uric acid.
4. The disintegration of fats sets free fatty acids. These are normally burned down to CO_2 and water, but if complete combustion does not occur, ketone bodies are accumulated. Thus

with certain unbalanced diets and in some pathologic conditions, acetylacetic and β-hydroxybutyric acid are accumulated.

5. Lactic acid is formed in the course of glycogen metabolism, but it is resynthesized into glycogen.
6. Oxalic acid is a product of the metabolism of some foods such as cocoa and spinach.

The formation of acids is a necessary step in some metabolic processes; therefore some acids should not be considered as waste products.

Base set free in the organism is derived mainly from vegetable foodstuffs which contain salts of organic acids. For example, in the metabolism of sodium citrate the citrate is oxidized and Na is left over, which with CO_2 and H_2O forms sodium bicarbonate. The urine of herbivorous animals is alkaline, because there is an excess of base (K, Na, Ca, and Mg) in the final products of the metabolism of vegetable foodstuffs. When these animals are kept fasting, as in Claude Bernard's famous experiment, the urine becomes acid.

The two factors of acid-base equilibrium—pH and alkali reserve—are kept within the normal limits by the excretion of the excess acid or base. There are several mechanisms for this, but undoubtedly the body is better equipped for the neutralization and elimination of acid than for that of base.

Secretions must be divided into two groups: (*a*) those in which the product of secretion is normally reabsorbed (saliva, gastric juice, bile, pancreatic juice, etc.); (*b*) those in which the product of secretion is eliminated from the organism, as in the case of sweat. In normal conditions the acid-base balance is not finally dependent on secretions of the first group, but in pathologic conditions they may be the cause of serious disturbance; *e.g.*, repeated vomiting leads to loss of Cl^-, and the relative increase in Na^+ leads to alkalosis. A permanent biliary or pancreatic fistula results in loss of base, and if adequate precautions are not taken acidosis results.

There are five mechanisms that take part in the regulation of the acid-base equilibrium: the blood, the respiratory, renal, and intestinal mechanisms, and finally the tissues.

Mechanisms in the blood. These mechanisms have already been discussed. They consist of (*a*) buffer systems in the plasma; (*b*) hemoglobin; (*c*) the selective permeability of the red cell membrane (see Chap. 30).

Respiratory mechanism. The excitability of the respiratory center plays an important part in the regulation of the acid-base equilibrium, because an increase in pulmonary ventilation increases the elimination of CO_2, and a decrease has the opposite effect, thus maintaining constant within the normal limits the ratio of free CO_2 to combined CO_2, as expressed in Henderson's equation. The respiratory center is so sensitive that a rise of 1 mm. in CO_2 partial pressure causes a 60 per cent increase in pulmonary ventilation. Drugs that depress the respiratory center (*e.g.*, morphine) cause CO_2 retention. The respiratory regulation of the acid-base balance is closely related to the regulation of respiratory movements, which will be fully discussed in Chaps. 32 and 33, particularly under "Anoxia."

Renal mechanism. The lung is the principal route for the elimination of CO_2, and the kidneys for the elimination of fixed acid and base. Urine secreted by the kidney can vary considerably in pH; according to the physiologic circumstances the pH varies between 7.4 and 4.8. The kidney regulates the elimination of acids by the selective excretion of acids and by the formation of ammonia (see page 741).

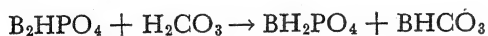
Selective elimination of acids, with the formation of acid urine, occurs when acid waste products are formed in the course of normal or abnormal (ketone bodies) metabolism, and when fixed acids (sulfuric, chlorhydric, phosphoric) or acidogenous salts (ammonium chloride, $CaCl_2$) are ingested. The kidney has a limited capacity for concentrating acids (the urinary pH cannot be lower than 4.8). If larger amounts of fixed acids must be eliminated they must be accompanied by a proportionate quantity of base taken from the alkali reserve, which is thus depleted.

For example, if a subject drinks a certain quantity of a dilute solution of HCl (concentrated solutions are caustic) the acid is eliminated through the kidney, not simply as HCl, but neutralized; otherwise the urinary pH would fall below the maximum limit of 4.8. If the base used for this purpose is Na taken from plasma bicarbonate, the alkali reserve is diminished, but the kidney can form ammonia, which is used instead of fixed base; thus the alkali reserve is spared.

Buffer systems in urine. Any solution containing a weak acid and its salt is a buffer system. The urine contains several of these systems, of which the best known is the acid phosphate–basic phosphate system. Henderson's equation for this system is the following:

$$cH^+ = K_1 \frac{BH_2PO_4}{B_2HPO_4}$$

This system is also found in the blood, and if the urine is acid, as it usually is in man, the numerator increases and the denominator decreases. Acid phosphate is formed by the following process:

$$B_2HPO_4 + H_2CO_3 \rightarrow BH_2PO_4 + BHCO_3$$

The bicarbonate thus formed is retained in the organism and goes into the alkali reserve.

If this were the only renal mechanism for controlling the acid-base balance, at least part of the fixed base would be lost; especially Na and K would be excreted as acid phosphates. Another mechanism for sparing base is the formation of ammonia, which takes the place of Na and K in the neutralization of acid ions eliminated in the urine (see pages 742–743).

Usually, in man, urinary ammonia accounts for approximately 4 per cent of the total urinary N, but when more acid is formed or ingested, the urine not only becomes more acid, but ammonia content also rises and can account for as much as 20 per cent of total urinary N.

To assess the part played by the kidney in the elimination of acids, total urinary acidity (determined by titration and expressed in cubic centimeters of 0.1 *N* acid solution) and urinary ammonia (expressed as cubic centimeters of 0.1 *N* ammonia) should also be estimated. The sum of these factors gives the urinary acid excretion index, which is an indirect way of determining the degree of acidity of the organism. Of course, if there are infectious processes in the bladder caused by microorganisms capable of fermenting the urine and producing NH_3 at the expense of urea, this index loses significance.

Origin of urinary ammonia. Ammonia excreted in response to an excess formation of acid is formed in the kidney. The following facts are evidence of this: (*a*) the amount of NH_3 in blood (0.46 mg. per liter) is not sufficient to account for the large quantities of NH_3 secreted in the urine in acidosis; (*b*) experimental total nephrectomy does not cause an increase of blood NH_3; (*c*) venous blood leaving the kidney has a higher NH_3 concentration than arterial blood entering the kidney.

At one time it was thought that ammonia was formed in the kidney from urea in the blood. Van Slyke and his associates[1] have shown that in the dog the greater part is formed from glutamine. Acidosis was provoked by the ingestion of HCl, and a difference in the glutamine in renal arterial and venous blood was found, which could not be accounted for by excretion of glutamine, but which corresponded to the increase in urinary ammonia. Alkalosis provoked by the ingestion of sodium bicarbonate, on the contrary, caused a decrease in the glutamine removed by the kidney from the blood. A glutaminase which catalyzes the formation of NH_3 from glutamine has been found in renal tissue. Van Slyke suggests that, beside this main source of ammonia, amino acids also contribute to its formation.

Renal control of alkalosis. When there is an excess of alkali in the blood, as after the ingestion of bicarbonate or citrate or a vegetable diet, the urine becomes alkaline and the pH can rise as far as 7.8. Urinary ammonia diminishes and can even totally disappear. Phosphate is excreted as basic phosphate, and part of the cations are eliminated as bicarbonates, as can be seen by adding a strong acid to the urine, in which case CO_2 is released and bubbling is observed. Another renal mechanism that becomes active in alkalosis is the selective elimination of citric acid, which can be increased a hundred times after ingestion of sodium bicarbonate. There is some evidence that citric acid is formed in the kidney under these conditions.[2]

Intestinal mechanism. The part played by the intestine in the regulation of the acid-base balance is not yet well known. Certain cations, such as Ca^{++} and Mg^{++}, are excreted by the intestine, but this does not seem to be a selective excretion conditioned by the acid-base balance. For example, if an excess of $CaCl_2$ is given, Ca is eliminated in great part through the gut, and Cl^- is retained. The anion in excess diverts Na from bicarbonate and the alkali reserve dimin-

[1] Van Slyke, D. D., *et al.*, *J. Biol. Chem.*, **150**, 481, 1943.

[2] Orten, J. M., and A. H. Smith, *J. Biol. Chem.*, **128**, 101, 1939.

ishes; thus in this case ($CaCl_2$ acidosis) the intestine not only does not contribute to maintain the acid-base balance but is the cause of its disturbance by the selective excretion of Ca^{++}.

Mechanisms in the tissues. The tissues undoubtedly play a part in the regulation of the acid-base equilibrium, but the mechanism of their contribution is not yet well known. The factors that concur in the acid-base equilibrium (pH, alkali reserve) are usually estimated in the blood, which is only 7 per cent of the total body weight, and probably there are equilibriums between tissue cells and fluids similar to those existing between erythrocytes and plasma. There is evidence that the tissues are important in the regulation of acid-base balance. For example, asphyxia or the ingestion of HCl causes a decrease in the alkali reserve. In the latter case the decrease in alkali is due to acid intoxication; in the former, to the shift of base from the blood. In asphyxia, breathing mixtures of CO_2 and O_2 rapidly reestablishes the alkali reserve and improves the condition of the patient; in acid intoxication, on the contrary, this would be a dangerous procedure. According to Yandell Henderson, the displacement of base that occurs in asphyxia should not be considered as acidosis.

Other mechanisms. There are perhaps other mechanisms for regulating the acid-base balance, but nothing definite is known about them. The excretion of sweat does not play an active part, because there is no selective elimination of cations or anions through the sweat glands as through the kidneys. The adrenal glands play an important part, which will be considered when dealing with these glands, in maintaining the balance of Na and K.

METHODS FOR DETERMINING THE STATE OF THE ACID-BASE EQUILIBRIUM

These methods will be only summarily considered here.

In the blood the pH is commonly determined by the potentiometric method; less frequently, by the indicator method. The pH and the alkali reserve (CO_2 combining power of plasma) are the two factors that must be known to establish the condition of the acid-base equilibrium. The gasometric method for the determination of the alkali reserve is much simpler than the volumetric method, which uses titration of bicarbonate by adding an excess of acid and then titration of the acid with NaOH, using phenol red as indicator.

In alveolar air the percentage of CO_2 can be determined and the partial pressure calculated. The CO_2 tension of alveolar air is the same as in the blood and one of the factors in Henderson's equation is thus known. The normal concentration of CO_2 in alveolar air is 4.7 to 6.8 volumes per cent. The method is subject to error due to variations in alveolar CO_2 concentration in the lung, caused by changes in pulmonary ventilation.

In urine the acid excretion index is obtained by adding the total acidity (in cc. 0.1 *N* acid) to the urinary ammonia (in cc. 0.1 *N* NH_3) calculated in the urine excreted during 24 hr. This is an excellent way of determining the acid-base balance of the body when the kidneys are functioning normally. In a healthy adult this index varies from 0 to 1,600; higher values are a sign of excess acid production. In nephritis this method cannot be used because in this condition there can be a decrease in the alkali reserve, without a corresponding increase in urinary excretion of acid.

Bicarbonate tolerance is also an indirect method. Sodium bicarbonate is given by mouth until the urine becomes alkaline. In normal subjects 5 to 10 gm. is needed. There is no acidosis if the urine becomes alkaline with less than 0.5 gm. per kg. of body weight.

BIBLIOGRAPHY

AUSTIN, J. H., and G. E. CULLEN, "Hydrogen Ion Concentration of the Blood in Health and Disease," Williams & Wilkins, Baltimore, 1926.

VAN SLYKE, D. D., Studies of Acidosis. The Normal and Abnormal Variations in the Acid-base Balance of the Blood, *J. Biol. Chem.*, **48**, 153, 1921.

VARELA-FUENTES, B., "Acidosis y Alcalosis en la Clínica," 2d ed., El Ateneo, Buenos Aires, 1941.

CHAPTER 32

Regulation of Respiration

THE DEPTH AND FREQUENCY of breathing, and therefore the pulmonary ventilation, are adapted in normal circumstances to the needs of the body by special regulatory mechanisms.

The respiratory muscles are under voluntary control, *i.e.*, their activity can be modified by impulses originated in the cerebral cortex, but this cortical control is limited. Thus breathing can be stopped voluntarily for a certain time, but not long enough to produce death by asphyxia. The respiratory rhythm can be modified for the emission of speech or in singing; also in strong emotional states. Nevertheless impulses from the cerebral cortex are not necessary for breathing, which continues during sleep or anesthesia and after removal of the whole cortex. The most important neural mechanism for the control of respiration is situated in the medulla. It is called the respiratory center, and it is under the influence of nervous and humoral factors.

The respiratory center has multiple effects. In the first place it coordinates the activity of the respiratory muscles so as to produce inspiration and expiration. Even the strongest voluntary effort cannot provoke inspiratory movements in one half of the thorax and expiratory movements in the other half; all the muscles taking part in inspiration (thoracic, laryngeal, dilators of the nares, and perhaps the smooth bronchial muscles) contract in a coordinated sequence. The respiratory center, under the influence of nervous and humoral factors, also regulates the frequency and strength of contraction of the respiratory muscles, and in special circumstances brings into activity the accessory respiratory muscles, thus controlling the rate and depth of breathing. This is known as the regulation of respiration.

The following aspects of the mechanism that controls breathing will be considered here: (*a*) the respiratory centers; (*b*) factors that condition the automatic rhythmic activity of the respiratory centers; (*c*) factors that regulate pulmonary ventilation.

THE RESPIRATORY CENTERS

Legallois in 1810, and Flourens in 1842, located a small region in the floor of the fourth ventricle, the destruction of which caused death due to the stopping of respiration. Flourens named this region the "*noeud vital*" (vital knot), a term which should be abandoned because it is incorrect, but which draws attention to the importance of this area for the maintenance of life.

The location of the respiratory centers has been established by experiments in which they have been destroyed or stimulated.

If the brain is sectioned in successive horizontal planes, no changes in breathing will be observed until a section has been made at the level of the rostral end of the pons and immediately caudal to the inferior colliculi. All sections above this level permit apparently normal breathing and normal respiratory responses to most nervous and chemical stimuli. Nevertheless some disturbances exist, as a section through the mid-part of the colliculi suppresses thermic polypnea (increased respiratory rate provoked by heating); and a section slightly caudal to this one suppresses reflex polypnea in light anesthesia (Fig. 152).

Sections at lower levels produce increasingly severe alterations in breathing, until it ceases completely when the section is made through the medulla at the level of the apex of the calamus scriptorius. Localized destruction and stimulation with needle electrodes, so as to affect only limited areas, have shown the exact location of

the center, or centers, in the reticular formation of the medulla, caudal to the entrance of the acoustic nerve and dorsal to the inferior olivary nuclei.

Stimulation with fine electrodes has demonstrated the existence of certain areas that provoke only inspiratory movements and of others that cause only expiratory movements.[1]

The inspiratory and expiratory areas have been accurately located by means of many experiments with needle-electrode stimulation. The two areas partially overlap, but it has been possible to establish the limits of each. In the cat, according to Pitts and his associates, the expiratory neurons are placed dorsally to the inspiratory neurons, nearer the floor of the fourth ventricle and extending farther in a cephalic direction. The inspiratory neurons are situated near the olivary nuclei, extending farther caudally. In the monkey and the dog similar areas have been located.

Although the expiratory and inspiratory neurons mingle in parts, they are functionally separated. Each one of the areas has all its neurons in close synaptic connection and acts reciprocally with the other area; thus, in localized stimulation of the inspiratory neurons, excitation spreads over the whole area and is accompanied by inhibition of the expiratory area.

Neurons on each side of the mid-line are connected by commissural fibers. A longitudinal section along the mid-line does not modify the respiratory rhythm if the vagi are intact, but if one vagus nerve is cut the respiratory rate of the denervated side diminishes; if both vagi are cut, each side of the thorax moves with an independent rhythm.

Apart from the respiratory center in the medulla, another center (or perhaps several) has been located in the upper part of the spinal cord. Respiratory movements have been observed in decapitated animals (section of the nerve centers between the medulla and the spinal cord) in certain circumstances, *e.g.*, in newborn animals, in adult animals injected with strychnine (which increases the excitability of the nerve centers), in animals maintained alive for a long period by means of artificial respiration, and in some birds. These movements provoked by spinal respiratory centers are not adequate to maintain normal breathing. The spinal centers are normally subject to the predominating influence of the medullary centers, but it is possible that when the latter are depressed the spinal centers play a more prominent part, causing disturbances in rhythm, such as periodic breathing (see Chap. 33).

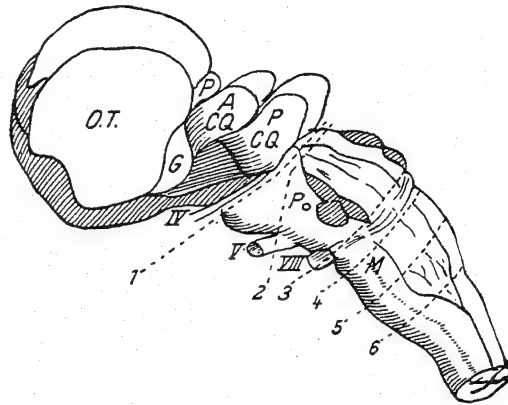



FIG. 152. Diagram of brain stem showing levels at which cross sections have been made to localize the respiratory center. Between 1 and 2 the pneumotaxic center is located; the respiratory center is situated between 4 and 5. *OT*, thalamus; *P*, pineal; *G*, geniculate body; *ACQ*, anterior colliculi; *PCQ*, posterior colliculi; *Po*, pons; *M*, medulla; IV, fourth nerve; V, fifth nerve; VIII, eighth nerve. (*After Lumsden, T., J. Physiol., vol. 57, p. 153, 1923.*)

The reciprocal innervation of respiratory muscles has been demonstrated in the same way as it has in the flexor and extensor muscles of the limbs. Contraction of the inspiratory muscles is accompanied by inhibition of the expiratory muscles, and vice versa.

Rhythmic, alternating activity of the inspiratory and expiratory centers is under the influence of impulses arising in the lungs and in the pneumotaxic center (see further on). Rhythmicity of the impulses sent out by the respiratory centers has been observed by recording the action potentials in nerves of respiratory muscles, especially the phrenic nerves and the nerves of the internal and external intercostals (Adrian). In quiet breathing the inspiratory center sends out series of impulses during inspiration, which are interrupted by periods of inactivity corresponding to expiration. These discharges are of low frequency, but their rate increases in more active breathing, *e.g.*, when the CO_2 concentration in arterial blood increases.

The expiratory center is almost inactive in quiet breathing, and sends out few if any im-

[1] PITTS, R. F., H. W. MAGOUN, and S. W. RANSON, *Am. J. Physiol.*, **126**, 673, 1939.

pulses. Expiration in this condition is therefore mainly passive and produced by the elastic recoil of the thorax. When breathing is more active, *e.g.*, in muscular exercise, the expiratory center also becomes active. It then sends out rhythmic impulses alternating with the inspiratory center, and the expiratory muscles contract.

Automatic Rhythmic Activity of the Respiratory Centers

The inspiratory and expiratory centers cannot by themselves maintain a normal respiratory rhythm, because the inspiratory center discharges impulses continuously and provokes sustained inspiration (*apneusis*). This activity of the inspiratory center is periodically inhibited by impulses from the pneumotaxic center and from the lungs, and expiration can thus take place.

Pneumotaxic center. This center is situated in the midbrain, immediately caudal to the colliculi (Fig. 152). It receives impulses from the hypothalamic centers which control body temperature. If connections between the hypothalamus and the pneumotaxic center are severed in the dog, heat polypnea is suppressed. The pneumotaxic center receives impulses from the inspiratory center and stimulates the expiratory center. The latter inhibits the inspiratory center (Gray). This circuit assures alternation of inspiratory and expiratory movements. It continues to function after both vagi have been cut.

The Hering-Breuer reflex. One of the most important mechanisms that automatically regulates the respiratory rhythm is a reflex arising in the lungs. It was discovered by Hering and Breuer in 1869, and was further studied by Head in 1889. The Hering-Breuer reflex takes place as follows: Distention of the lungs, as in inspiration, originates centripetal impulses toward the respiratory centers, which inhibit inspiration in normal breathing and stimulate expiration in forced breathing. Retraction of the lungs, as in expiration, originates centripetal impulses which result in inspiration. The afferent impulses travel along fibers of the vagus nerve, the neurons of which are in the ganglion nodosum.

The rabbit is well suited for the study of this reflex because it has a fairly long slip of the diaphragm at each side of the ensiform cartilage which can be freed so as to record its movements without damage to its innervation and circulation. At the end of expiration the diaphragm begins to contract, and at the end of inspiration it relaxes. If the lungs are artificially distended beyond the normal limits, there is prolonged inhibition of the diaphragm. Conversely, if the lungs collapse and more than the normal amount of air is expelled from them, there is a prolonged tonic contraction of the diaphragm. All the other respiratory muscles respond in a similar way, *i.e.*, by relaxation of the inspiratory and contraction of the expiratory muscles when the lung is distended, and by relaxation of the expiratory and contraction of the inspiratory muscles when the lung collapses.

The Hering-Breuer reflex is originated by the mechanical distention (or retraction) of the lung, and it takes place when the lung is distended by an inert gas, such as nitrogen or hydrogen; changes in oxygen or CO_2 content play no part in this reflex. In short, at each respiratory movement (inspiration and expiration) a Hering-Breuer reflex takes place, stimuli are originated in the lung that cause the centers to discharge impulses provoking the opposite movement.

Adrian[1] has studied the Hering-Breuer reflex in the cat by registering the action potentials of the afferent fibers of the vagus. Distention (inspiration) and retraction (expiration) of the lung provoke afferent discharges in the vagus. The receptors are probably situated in the alveolar ducts. They are not sensitive to changes in the partial pressure of the gases in the lung, but respond to changes in the tension of the lung tissues. The impulses act mainly on the neurons of the inspiratory center; those originated by distention of the lung inhibit these neurons and those originated by retraction stimulate them.

In normal breathing at rest (eupnea), impulses are originated only at the end of inspiration, and they inhibit the inspiratory center. Impulses that stimulate inspiration arise only when the lung retracts (expiration) more than in normal quiet breathing; these possibly become important in dyspnea.

Section of both vagi nerves in the neck causes serious disturbance in respiratory rhythm by suppressing the Hering-Breuer reflex mechanism. Breathing is slowed down, there are deep inspirations separated by long pauses during which the thorax is held in a position halfway

[1] Adrian, E. D., *J. Physiol.*, **79**, 332, 1933.

between inspiration and expiration. About one hour after the vagi have been cut the respiratory rate increases, the mechanism of this recovery being still unknown.

If the vagi are cut above the origin of the recurrent laryngeal nerves, most of the laryngeal muscles are paralyzed and the vocal chords do not separate on inspiration, thus forming an obstacle to the passage of air. The cough reflex also disappears, owing to the section of the afferent fibers, so there is no reflex defense for the elimination of foreign bodies which fall into the larynx. This is a serious disturbance in the dog and may lead to death by infection of the lungs. The importance of the vagus in respiration varies according to the species; in guinea pigs double vagotomy is rapidly followed by death due to asphyxia caused by spasmodic contraction of the smooth muscles of the small bronchi.

An irregular respiratory rhythm is maintained after the connections between the pneumotaxic center and the medulla have been cut, if at least one vagus nerve is intact. It is also maintained after both vagi have been cut if the paths between the midbrain and the medulla are intact. If the fibers between the midbrain and the medulla and one vagus are cut, inspiration is prolonged and respiratory frequency diminishes. If then the other vagus nerve is cut, the inspiratory center no longer receives inhibitory impulses; an inspiratory spasm therefore takes place, which stops breathing (*apneusis*), and the animal dies from asphyxia. Records of action potentials in the phrenic nerve show a continuous discharge of nerve impulses, which can be inhibited, together with the inspiratory spasm, by stimulating the central end of one of the cut vagi.

There are three main types of normal breathing:[1] (*a*) eupnea, (*b*) sighing, and (*c*) panting. There are also several types of abnormal rhythm, such as Cheyne-Stokes breathing.

In eupnea (the usual type of breathing) frequency, depth, and the expiratory base line remain constant. Changes in metabolic needs modify frequency or depth, but the regularity of breathing is maintained.

Sighing is superimposed on the basic rhythm. Eupnea and sighing are observed after double vagotomy if the connections between the pons and the medulla are intact.

Panting, *e.g.*, heat polypnea in the dog, is caused by a marked increase in frequency with a decrease in depth (rapid, shallow breathing). Section of the midbrain between the colliculi severs the paths from the hypothalamus to the pneumotaxic center and suppresses heat panting.

FACTORS THAT REGULATE PULMONARY VENTILATION

Changes in the depth and frequency of breathing cause changes in pulmonary ventilation, adapting the gaseous interchange in the lungs to the requirements of the organism. Several factors play a part in regulating pulmonary ventilation. The principal ones are the sensitiveness of the respiratory center to CO_2 and the afferent impulses that come from different parts of the body.

Apart from these physiologic factors, several drugs stimulate or inhibit respiration. A drug that increases pulmonary ventilation can produce its effect by one of two mechanisms: as a direct stimulant, or indirectly by increasing the sensitiveness of the respiratory center, *i.e.*, lowering its threshold. In the first case the effect of the drug is added to that of CO_2, the physiologic stimulant; in the second case a normal CO_2 concentration in the blood acts as a stronger stimulus owing to the greater excitability of the center. It is not possible to say which of these mechanisms is the one by which respiratory stimulants act. According to Y. Henderson,[1] respiratory activity "is always the product of at least two main factors: two factors that are most conveniently denominated as (1) the strength of the stimulus and (2) the sensitivity or excitability of the respiratory center. This conception requires that carbon dioxide shall always be regarded as the sole chemical stimulus to respiration, either directly or through pH. All the other substances and conditions that increase or decrease breathing are denied as stimuli. They must rather be regarded as acting entirely by altering—either raising or lowering—the sensitivity of the respiratory center to carbon dioxide. Under this conception they are held to be stimulants, not stimuli. It is a conception that is indeed artificial and arbitrary, but extremely useful as a method of calculating

[1] HOFF, H. E., and C. G. BRECKENRIDGE, Levels of Integration of Respiratory Patterns, *J. Neurophysiol.*, **15**, 47, 1952.

[1] HENDERSON, Y., "Adventures in Respiration," Williams & Wilkins, Baltimore, 1938, p. 372.

results in a wide range of respiratory problems." Gray, however, suggests another interpretation of the facts (see page 298).

Sensitivity of the respiratory center to CO_2. The respiratory center responds up to a certain point to several factors, such as temperature and oxygen partial pressure, but undoubtedly it is most sensitive to carbon dioxide. There are, of course, individual differences, but in certain subjects an increase by 0.2 per cent in CO_2 concentration in the alveolar air doubles pulmonary ventilation. A decrease of the same magnitude causes transitory apnea. This variation in CO_2 concentration corresponds to a change in CO_2 partial pressure of 1.5 mm. Hg. The minimum (threshold) variation in CO_2 partial pressure that modifies breathing in man is approximately 0.075 mm. Hg. CO_2 exerts its influence on the centers, as is proved by experiments in animals in which the vasosensory areas (see page 295) have been denervated. This can also be demonstrated by cross-circulation experiments, in which the head of a dog *A* receives its blood supply from a second dog *B*. If the latter breathes air in which the CO_2 concentration is high, pulmonary ventilation increases in dog *A*.

The effect of CO_2 increases with its concentration in the alveolar air up to a maximum at 9 per cent; at higher concentrations CO_2 depresses the respiratory center.

The pH of the blood depends on the ratio of free CO_2 to bicarbonate, as is shown in Henderson's and Hasselbalch's equations. The injection of acid increases pulmonary ventilation, so the problem of whether CO_2 acts directly as a stimulus of the respiratory center or by modifying the hydrogen ion concentration, which would be the direct stimulus, has been the subject of much discussion.

At first this question could not be answered. Some workers (Haldane, Winterstein, etc.) considered the hydrogen ion concentration to be the physiologic stimulus of the respiratory center; others (Collip, etc.) thought it was the HCO_2^-. In certain circumstances pulmonary ventilation and the acidity of the blood do not vary simultaneously. Gesell supposed that this was due to several factors; for instance, CO_2 is more diffusible than H^+, and acid products may arise in the course of the cellular metabolism of the respiratory center. If CO_2 concentration in the blood increases without a change in the free CO_2–bicarbonate ratio (Henderson's equation), the pH does not vary, but there may be an increase in acidity in the tissues because of the greater diffusibility of dissolved CO_2. This can be demonstrated in several ways; *e.g.*, the flowers of *Symphytum peregrinum*[1] are blue in a neutral or alkaline medium and red in acid medium. Submerged in acid water of pH 5 to 6, they remain blue, but in a solution of carbonic acid and sodium bicarbonate of pH 7.4, they change to red, because CO_2 diffuses rapidly into the cells and acidifies them.

Gesell measured the variations in the pH of the cells of the respiratory center by means of electrodes placed in the cerebrospinal fluid, thus obtaining a continuous record of the changes taking place; simultaneously he obtained a record of pH variations in the blood. He concluded that stimulation of the respiratory center depends on changes in acidity within the center, changes which in some circumstances do not coincide with those occurring in the blood. These experiments demonstrate the importance of intracellular changes but do not solve the problem of whether cH^+ or H_2CO_3 is the stimulus of the respiratory center.

More recent work seems to show that H_2CO_3 has a specific effect on the respiratory center, independent of its effect on cH^+. The following facts, mainly due to research done by Krogh, are the principal arguments that support this statement: (*a*) prolonged breathing of air with CO_2 provokes intense hyperpnea, while there are insignificant changes in blood pH; (*b*) ingestion of ammonium chloride, an acidogenous salt, acidifies the blood but causes little or no hyperpnea. For example, breathing air with 4 per cent CO_2 increased pulmonary ventilation by 10 liters per min., accompanied by a decrease in pH of 0.04 to 0.045; ingestion of ammonium chloride decreased blood pH by 0.08 and increased pulmonary ventilation by only 0.3 liters per min. On the other hand cases of acidosis with hyperpnea have been observed, in which there is acapnia, and hyperpnea persists even when the CO_2 partial pressure in alveolar air is only one-half or one-quarter the normal value. It is not possible to discard the suggestion that in these cases hyperpnea is due to a reflex originated in the chemoreceptors of the vasosensory areas.

Another striking demonstration of the specific effect of CO_2 has been given by perfusing the medulla with blood in which constant acidity

[1] JACOBS, M. H., *Am. J. Physiol.*, 53, 457, 1920.

is maintained by CO_2 or by HCl; in the first case pulmonary ventilation increases to 3,069 per cent, while in the second it increases to only 128 per cent (Hooker *et al.*). Comroe, by injecting different solutions into the medulla, was able to stimulate respiration with buffer systems of bicarbonate, but only in a few cases could he obtain similar results with solutions of lactic acid or HCl.

The weight of the evidence is therefore in favor of the idea that stimulation of the respiratory center is due to a specific or preponderant effect of carbonic acid acting within the center, and not to changes in hydrogen ion concentration conditioned by the ratio of dissolved CO_2 to fixed CO_2.

The part played by changes in oxygen tension. The respiratory center is very sensitive to lack of oxygen; anoxia of a certain intensity and duration depresses the center and, if it lasts for more than a few minutes, may produce irreversible damage. Nevertheless when the oxygen partial pressure decreases, at first there is an increase in pulmonary ventilation, which may be fairly prolonged if anoxia is not too intense. This is due to an increase in the excitability of the respiratory center caused by impulses arising in the chemical receptors of the vasosensory areas (see further on). A fall in oxygen partial pressure increases primarily the depth of breathing and, in a lesser degree, the respiratory rate. These changes are less marked than those produced by increase in CO_2, but they are sufficient to cause a significant increase in pulmonary ventilation.

There are considerable individual variations in the minimum (threshold) decrease in oxygen tension that can modify pulmonary ventilation. Expressed as percentages of oxygen at sea-level pressure, significant changes begin in most subjects when the inspired air contains only 13 per cent oxygen, but some normal individuals respond with increased breathing when oxygen in the air drops to 18 per cent.

The symptoms provoked by a decrease in oxygen tension become severer as anoxia increases. A drop from 21 to 18 per cent (159 to 137 mm. Hg oxygen partial pressure) does not cause any disturbance in respiration. Between 16 and 12 per cent (121 to 91 mm.) pulmonary ventilation increases, there is a quicker pulse rate, incoordinated movements appear, and muscular efficiency diminishes. In an atmosphere of 12 per cent oxygen, a candle does not burn. Between 14 and 9 per cent (106 to 68 mm.) the respiratory and pulse rates increase even more. There is periodic (Cheyne-Stokes) breathing, cyanosis, vomiting, asthenia, and fatigue. Errors are committed when doing simple arithmetical calculations, and there are changes in behavior, the subject sometimes reacting violently. Between 10 and 6 per cent (76 to 45 mm.) the subject is excited, and there is intense cyanosis. The electrocardiogram shows alterations; the PR and RT intervals decrease, and the amplitude of T diminishes. Blood pressure falls, syncope may occur, and sometimes the subject becomes comatose. Respiration is at first deep and then becomes shallow and frequent, ending in inspiratory spasm. A decrease in oxygen percentage (or its equivalent partial pressure) of this magnitude is tolerated for only a very short time. A drop to 5 or 3 per cent (38 to 23 mm.) is rapidly fatal.

Accidents observed in rarefied air are due to the decrease in oxygen partial pressure and not to the fall in total pressure. Paul Bert observed many years ago that the total pressure can fall considerably without causing accidents if the subject breathes pure oxygen. Further details will be given when considering mountain sickness (see Chap. 33).

Vasosensory respiratory reflexes. The importance of vasosensory areas in the control of circulation has already been discussed (see Chaps. 17 and 20). The main vasosensory areas are situated in the aortic arch, innervated by the cardioaortic nerve, and the carotid sinuses at the bifurcation of the common carotid, innervated by the sinus nerves of Hering. There are two types of receptors in these areas; one type is sensitive to changes in arterial blood pressure (pressoreceptors), the other to changes in the chemical composition of the blood (chemoreceptors). The pressoreceptors are found in the wall of the arteries, and the chemoreceptors in the aortic and carotid bodies.

Changes in arterial blood pressure and in the chemical composition of the blood stimulate these receptors and originate impulses which travel along the cardioaortic and sinus nerves. These impulses modify the excitability of the respiratory centers and cause respiratory reflexes which are suppressed by total denervation of these areas. The part played by the vasosensory areas in the regulation of breathing has been

studied in experiments of total or partial denervation of the areas, destruction of the aortic and carotid bodies, perfusion of the carotid sinus with blood of controlled chemical composition at controlled pressure before and after cutting the sinus nerve or extirpating the carotid body, etc. A long series of important experiments, many of them by Heymans and his associates, have accumulated many facts confirming some classical ideas and adding new ones on the regulation of circulation and respiration.

The influence of pressoreceptors on respiration. A sudden increase in arterial blood pressure inhibits respiration and may cause transitory apnea. This is commonly seen when injecting adrenaline into dogs. Hypotension, such as is caused by inhalation of amyl nitrite, accelerates breathing. The reflex nature of these responses is demonstrated by the following experiment: The carotid sinus of a dog is perfused by anastomosis of the common carotid with the artery of a second dog. The blood pressure is increased or decreased in the second dog, and changes in respiratory rate are observed in the perfused animal. Denervation of the perfused carotid sinus suppresses the effect of variations in blood pressure on respiration.

Pressoreceptors are more important for the reflex control of circulation than of respiration. Thus section of both cardioaortic and sinus nerves causes prolonged permanent hypertension, but only transitory hyperpnea, and in some experiments no significant change in breathing has been observed. Pressor changes in the sinus provoke reflex changes in arterial blood pressure, which are proportionate to the sinus variation and last as long as the sinus pressure does not alter. The respiratory response, on the other hand, is not sustained; there is only transitory apnea or hyperpnea when the sinus pressure suddenly rises or falls.

The influence of chemoreceptors on respiration. Receptors in the aortic and carotid bodies which respond to changes in the chemical composition of the blood play a relatively small part in the regulation of circulation, but they are of considerable importance in the control of respiration. These chemoreceptors are stimulated by a decrease in oxygen tension, by an increase in hydrogen ion concentration or carbonic acid, and by nicotine, cyanide, H_2S, and other drugs.

The principal physiologic stimulus is the decrease in the oxygen partial pressure of the blood. Sensitiveness of chemoreceptors to small changes in oxygen partial pressure can be demonstrated in animals if care is taken to avoid depression by anesthetics. Decerebrated cats are 100 times more sensitive than cats anesthetized with chloralose, and 10,000 times more sensitive than dogs under the same anesthetic.[1] Records of action potentials in Hering's nerve show that chemoreceptors begin to discharge when oxyhemoglobin is 96 per cent saturated. Activity increases as oxygen saturation of hemoglobin diminishes, almost in linear relation, with a maximum at 15 per cent oxyhemoglobin. The decrease in oxygen partial pressure acts as a stimulus; anemic anoxia does not stimulate the chemoreceptors. Owing to their high sensitiveness to decreases in oxygen partial pressures, chemoreceptors should be considered as part of the mechanism controlling oxygen equilibrium in arterial blood in normal conditions, and not simply as an emergency mechanism, as experiments on anesthetized animals seemed to indicate.

Sensitiveness to an increase in hydrogen ion concentration or CO_2 has been demonstrated in perfusion experiments in which the carotid body and sinus nerves are intact. In physiologic conditions these stimuli are not of much importance, since a change of 0.1 in pH is needed to produce the minimum (threshold) response and variations of this magnitude seldom occur in normal subjects.

The chemoreceptors are not very sensitive to changes in arterial CO_2 partial pressure, which must be of the order of 10 mm. to produce effects. The respiratory centers, on the other hand, respond to changes in arterial CO_2 partial pressure of 0.075 mm. Hg. Therefore CO_2 cannot be considered as the normal stimulus of the chemoreceptors.

A salient feature of the chemoreceptors is their great resistance compared with that of the respiratory centers. Thus anoxia of a certain intensity and duration depresses the respiratory center, but does not disturb the chemoreceptors. This is well demonstrated by perfusing the carotid sinus with fluid free from oxygen. Hyperpnea is provoked and sustained while the perfusion lasts; it ceases immediately on adding oxygen to the perfusion fluid. Resistance of the chemoreceptors to lack of oxygen is perhaps an

[1] Alvarez-Buylla, R., *Arch. Inst. cardiol. México,* **21,** 724, 1951.

important factor in maintaining respiration in serious anoxia. This chemical reflex mechanism is the *ultimum moriens* in the control of respiration (Comroe and Schmidt).

Chemoreceptors are also sensitive to other stimuli, *e.g.*, changes in temperature, but as these must be of great magnitude to reach the threshold, they are not of physiologic importance. Of more importance is the effect of drugs. Nicotine, lobeline, cyanides, and H_2S in adequate doses stimulate the chemoreceptors and produce an increase in pulmonary ventilation. The following facts are evidence of this statement:

1. Complete denervation of the receptors by section of both cardioaortic and sinus nerves suppresses the respiratory effects of these drugs.
2. Perfusion of the carotid sinus of a dog with fluid containing one of these drugs has the same effect as the injection of the drug into the general circulation.
3. Very small doses of the drugs, which have no effect when injected into the general circulation, produce a respiratory response when injected into the aorta, thus reaching the chemoreceptors in higher concentration.
4. Injection of the drugs provokes the appearance of action potentials in the sensory nerves of the receptors, showing that these have been stimulated.

Hyperpnea in muscular exercise. Muscular exercise causes an increase in pulmonary ventilation, which has been attributed to the effect of CO_2, owing to the exquisite sensitivity of the respiratory center to minute increases in arterial CO_2 tension. Evidence recently obtained shows that other factors are even more important.

Barcroft and Margaria observed that the maximum pulmonary ventilation that could be provoked by breathing air mixtures with CO_2 was of 50 to 60 liters per minute. On the other hand, violent muscular exercise could increase pulmonary ventilation to twice this figure, although the sensation of distress in breathing was not so marked as when breathing the CO_2 mixtures.

In muscular exercise, changes in alveolar CO_2 tension are not parallel to changes in pulmonary ventilation. CO_2 tension increases in moderate exercise and usually diminishes as the intensity of the exercise increases, being generally below normal in violent exercise. Moreover, hyperventilation increases before there is any significant increase in CO_2 output.

Lactic acid, released in the course of muscular contraction, has been considered the additional factor necessary to produce this hyperpnea; but in moderate exercise there is a threefold to fourfold increase in pulmonary ventilation without any significant increase in blood lactic acid.

It is generally accepted that hyperpnea observed in exercise is due to reflexes arising in the muscles, which increase the excitability of the respiratory center. There is some experimental evidence in support of this assertion. Electrical stimulation of the muscles, or even passive movement of a limb, increases pulmonary ventilation in the dog. Occlusion of the blood vessels of the limb does not suppress this response, but denervation or anesthesia of the nerves with cocaine abolishes it. Whether or not metabolic products originated by muscular contraction act locally is still under discussion.

Reflexes originated in the large veins and auricles, similar to the reflex acceleration of the heart rate provoked by distention of these structures (Bainbridge's reflex), have also been considered factors in the mechanism of the hyperpnea of muscular exercise. Hyperthermia stimulates breathing, but it cannot be a factor in the initial hyperventilation of exercise, because the rise in temperature occurs much later. Perhaps adrenaline, which stimulates respiration in doses too small to cause a rise in blood pressure, also concurs in the production of this hyperpnea.

In summary, experimental evidence has shown that hyperpnea of exercise is caused mainly by a reflex increase in the sensitivity of the respiratory center. An increase in CO_2 or lactic acid during the first stages is of secondary importance and may not take place although hyperpnea occurs.

Other factors that act on respiration. Pi-Suñer has reported a specific sensitivity of the lung to a local increase in CO_2, but there is no definite proof that this mechanism acts within the normal limits of variation of CO_2 partial pressure.

Stimulation of the trigeminal nerve endings in the nasal mucosa by inhalation of an irritant gas (Cl_2 or NH_3) stops respiration. Initial apnea in ether anesthesia is due to the same cause; it ceases when sufficient CO_2 has accumulated to overcome the inhibitory reflex impulses. As anesthesia progresses the nerve centers lose their sensitivity and the apneic pauses disappear.

Irritation of the nasal mucosa with weaker stimuli provokes sneezing.

Coughing is a reflex originated in the pharynx (innervated by the glossopharyngeal), larynx, trachea, or bronchi (innervated by the vagus). The most sensitive area is situated in the larynx above the vocal chords. Vagotomy above the recurrent laryngeal nerves suppresses the coughing reflex by anesthetizing the mucosa where it originates. In the dog this defensive reaction is of considerable importance in avoiding the entrance of foreign bodies into the lungs; hence the seriousness of vagotomy in this species.

Coughing consists of a deep inspiration followed by a sudden sharp expiratory movement, keeping the glottis closed until it is overcome by the pressure in the lung. Thus a strong current of air is originated, which carries away the stimulating agent (mucus, foreign bodies, etc.). Sneezing is caused by a similar mechanism. Coughing can be performed voluntarily, but once deep inspiration has started as a result of a stimulus of sufficient strength, it is very difficult to inhibit it voluntarily; perhaps the Hering-Breuer reflex plays a part in the expiratory phase of coughing. Coughing can be diminished in bronchitis by shallow breathing; thus the probability of a strong expiratory movement, which may follow a deep inspiration, is reduced.

Anesthesia and other conditions that depress the nerve centers also depress the cough reflex and increase the danger of pulmonary infection due to retention of secretions and foreign bodies.

Intense irritation of the larynx provokes spasm of the glottis, which can be considerably dangerous in children.

Deglutition reflexly inhibits respiration.

Painful sensations and emotions (fear, rage, etc.) increase the respiratory rate and pulmonary ventilation. The influence of the cortex on respiration is also evident in the spasmodically interrupted breathing of sobbing and laughter, but the mechanism of this cortical effect is still unknown.

Respiratory rhythm, frequency, and depth are the result of the integration of all the factors that concur in the adaptation of pulmonary ventilation to the needs of the organism. As Comroe[1] has very aptly expressed it, "respiration is controlled not by reflexes alone, not by chemical stimulation of the medulla alone, but by the proper interaction of both factors. No reflex, no matter how strong, can stimulate respiration if the arterial CO_2 tension has been lowered abnormally; no chemical stimulant, no matter how great, can produce rhythmic breathing if the medullary centers have been completely cut off from all nervous influences."

[1] Comroe, J. H., *Physiol. Rev.*, 24, 333, 1944.

The multiple-factor theory of the control of respiratory ventilation. Respiration is controlled by complex mechanisms and, according to Gray, there is no single "true" stimulus. Changes in the pH and partial pressures of O_2 and CO_2 in blood vary according to the cause which has provoked hyperpnea (Table 29), a fact that has led Gray to propose a multiple-factor theory of respiratory control.

Table 29. Maximum Pulmonary Ventilation in Four Types of Hyperpnea

Condition	*Maximum ventilation, liters/min.*	*Changes in arterial blood*		
		P.p. O_2	*P.p. CO_2*	*pH*
Anoxia	12	↓	↓	↑
Breathing CO_2	70	↑	↑	↓
Metabolic acidosis	35	↑	↓	↓
Exercise (moderate)	50	0	0	0
Exercise (violent)	100	0	↓	↓

Source: Gray, J. S., "Pulmonary Ventilation and Its Physiological Regulation," Charles C Thomas, Springfield, Ill., 1950.

Functional integration of the respiratory centers (inspiratory and expiratory) with the pneumotaxic center and the Hering-Breuer reflexes has to do with the control of periodicity and depth of breathing, but not with direct control of the respiratory minute volume. Other factors also take part in the regulation of pulmonary ventilation, though finally the same centers and paths come into play.

Oxygen and CO_2 tensions and the pH in the blood can vary in different ways, according to the cause of hyperventilation; therefore there cannot be only one factor of adaptation. The multiple-factor theory of Gray is based on three main principles:

1. The multiple-factor principle states that a number of factors exert independent effects upon respiratory ventilation. These factors are (*a*) changes in the pH and CO_2 acting on central (medullary) chemoreceptors; (*b*) changes in O_2 tension acting on peripheral

chemoreceptors; (*c*) afferent impulses arising in the course of muscular exercise; (*d*) pressor-receptor reflexes arising in the large arteries and veins; (*e*) thermic reflexes arising in thermoreceptors in the skin sensitive to changes in the temperature of the environment, and in the hypothalamus sensitive to changes in the temperature of the blood; (*f*) afferent impulses from pain receptors and psychic influences. The most important are the chemical changes (pH and CO_2 and O_2 tensions) and muscular reflexes.

2. The interdependence principle states that a change in any one of the respiratory factors usually brings about changes in one or more of the other factors.
3. The algebraic-summation principle states that the actual ventilation is defined as the algebraic sum of the partial effects of the separate agents.

BIBLIOGRAPHY

BARCROFT, J., and R. MARGARIA, Some Effects of Carbonic Acid on the Character of Human Respiration, *J. Physiol.*, 72, 175, 1931.

BAYER, G., Regulation der Atmung, in Bethe's "Handbuch der normalen und pathologischen Physiologie," 2, Springer, Berlin, 1925.

COMROE, J. H., The Hyperpnea of Muscular Exercise, *Physiol. Rev.*, 24, 319, 1944.

COMROE, J. H., and C. F. SCHMIDT, The Part Played by Reflexes from the Carotid Body in the Chemical Regulation of Respiration in the Dog, *Am. J. Physiol.*, 121, 75, 1938 (see also *Physiol. Rev.*, 20, 115, 1940; *Ann. Rev. Physiol.*, 3, 151, 1941).

CORDIER, D., and C. HEYMANS, "Le Centre respiratoire," Hermann & Cie, Paris, 1935.

GESELL, R., Regulation der Atmung und des Kreislaufes, *Ergebn. d. Physiol.*, 28, 340 and 692, 1929.

GESELL, R., *Ergebn. d. Physiol.*, 43, 477, 1940 (Functions of the respiratory center).

GRAY, J. S., The Multiple Factor Theory of the Control of Respiratory Ventilation, *Science*, 103, 739, 1946.

GRAY, J. S., "Pulmonary Ventilation and Its Physiological Regulation," Charles C Thomas, Springfield, Ill., 1950.

HEYMANS, C., and J. J. BOUCKAERT, Dissociation des deux sensibilités réflexogènes de la bifurcation carotidienne: sensibilité chimique et sensibilité à la pression, *Compt. rend. Soc. de biol.*, 112, 1240, 1933.

HEYMANS, C., J. J. BOUCKAERT, and L. DAUTREBANDE, Sinus carotidien et réflexes respiratoires, *Arch. internat. de pharmacodyn. et de thérap.* 39, 400, 1930; 40, 54, 1931.

PI-SUÑER, A., The Regulation of the Respiratory Movements by Peripheral Chemo-receptors, *Physiol. Rev.*, 27, 1, 1947.

PITTS, R. F., Organization of the Respiratory Center, *Physiol. Rev.*, 26, 609, 1946.

CHAPTER 33

Disturbances in Respiratory Function

PULMONARY RESPIRATION is usually automatic and does not intrude into the conscious sphere; but in contrast to other visceral functions, it can enter consciousness if the attention is directed toward it, and it can, to a certain extent, be modified at will. Normal automatic respiration is called eupnea.

DYSPNEA

Dyspnea literally means difficult breathing. It involves an objective factor, the active participation of the accessory inspiratory and expiratory muscles, and a subjective factor, the sensation of effort in breathing. According to Meakins the second factor is the more significant; he defines dyspnea as "the consciousness of the necessity for increased respiratory effort." The subjective factor, however, is absent in certain cases of dyspnea, *e.g.*, in anesthetized or comatose patients.

"Hyperpnea" means an increase in pulmonary ventilation, and "tachypnea," increased respiratory frequency. These conditions should be distinguished from dyspnea, although they can be found in dyspneic subjects, because they can exist without dyspnea, *e.g.*, during moderate exercise.

The change from eupnea to dyspnea takes place gradually. Usually there is a close connection between vital capacity and the ease with which dyspnea occurs. As the vital capacity diminishes, liability to dyspnea increases, but vital capacity is not the only factor conditioning dyspnea.

Dyspnea can be the result of many causes, but there are three main ones: (*a*) oxygen insufficiency; (*b*) CO_2 excess; (*c*) direct or reflex hyperexcitability of the respiratory center. Usually the three are combined. The principal functional disturbances that provoke dyspnea are (*a*) insufficient oxygenation of the blood as it passes through the lungs; (*b*) acidosis; (*c*) increased metabolic rate; (*d*) lesions in the vicinity of the respiratory center; (*e*) emotional and nervous excitation; (*f*) hyperexcitability of the Hering-Breuer reflex.

Insufficient oxygenation of the blood.[1] Insufficient oxygenation of the blood stimulates the respiratory center by reflexes arising in the chemoreceptors in the aortic and carotid bodies. Lack of oxygen in the blood may be due to (*a*) disturbances in the lung; (*b*) low oxygen partial pressure in the inspired air; (*c*) circulatory disturbances; (*d*) deficiency in the oxygen-carrying capacity of the blood due to lack or alteration of hemoglobin.

Disturbances in the lung. Gaseous interchange in the lungs may be hindered by obstruction of the larynx, trachea, or bronchi; by spasm of the smooth muscles of the bronchi (asthma); by the presence of fluid in the alveoli (alveolitis, edema, pneumonia); by a decrease in the respiratory surface caused by pulmonary collapse (pneumothorax), compression of the lung (pleural effusion), inflammatory processes (pneumonia), obstruction of a pulmonary blood vessel (infarct), etc.; by thickening of the pulmonary epithelium; by loss of elasticity of the pulmonary tissue (emphysema); by partial paralysis of the respiratory muscles. When the respiratory sur-

[1] The causes of insufficient oxygenation of the blood will be studied in detail in the section on anoxia. In order to avoid repetition they will only be mentioned here.

face is diminished and in cases of emphysema, the Hering-Breuer reflex is also exaggerated and rapid shallow breathing (tachypnea) results.

Low oxygen partial pressure in the inspired air. This occurs at high altitudes and in mines and submarines; also in the course of anesthesia with gases of relatively weak effect such as nitrous oxide and ethylene, which must therefore be given in concentrations of 80 per cent and even more. These gases should be given mixed with pure oxygen, so as to maintain an adequate oxygen partial pressure.

Circulatory disturbances. Among these the following can be mentioned: short circuits within the pulmonary circulation or the heart (congenital malformation); pulmonary congestion due to heart insufficiency; sclerosis of the pulmonary arteries; and circulatory collapse (which also causes retention of CO_2).

Deficiency in the oxygen-carrying capacity of the blood. This occurs in cases of severe anemia, or when hemoglobin is transformed into carboxyhemoglobin (carbon monoxide intoxication) or into methemoglobin or sulfhemoglobin.

Acidosis. Acidosis is due to the increase of free CO_2 in relation to fixed CO_2 (bicarbonate), and this results in stimulation of the respiratory center. This cause of dyspnea is frequently combined with others. Acidotic dyspnea is observed after inhalation of CO_2; after ingestion of acidogenous salts, such as calcium or ammonium chloride; and in endogenous uncompensated acidosis (diabetic coma, uremia).

Increased metabolic rate. Exercise does not provoke dyspnea in a normal subject unless it is fairly strenuous. The important factor in this case seems to be reflex stimulation of the respiratory center by impulses arising in the active muscles; lack of oxygen and accumulation of CO_2 apparently play a secondary part. Moderate exercise causes dyspnea in subjects with certain chronic cardiac diseases (cardiac dyspnea).

Lesions in the respiratory center. Many disturbances of the brain centers (encephalitis, brain tumors, cerebral hemorrhage, edema, etc.) provoke dyspnea. Intracranial pressure is increased in these cases and probably there is an insufficient blood supply to the respiratory center, which causes local acidosis, this being the main factor in the abnormal stimulation of the center. Certain drugs (caffeine, strychnine, atropine, etc.), acting directly on the center, also provoke dyspnea.

Emotional and nervous excitation. Psychically unstable individuals (neurotics, hysterical persons) sometimes have attacks of dyspnea, similar to those accompanying strong emotion. In these cases cortical excitation probably spreads down to the respiratory center.

Hyperexcitability of the Hering-Breuer reflex. Reflex dyspnea. In certain cases the Hering-Breuer reflex is abnormally strong, and causes rapid shallow breathing and consequently insufficient pulmonary ventilation, thus providing another cause of dyspnea. An exaggerated Hering-Breuer reflex is observed when the respiratory surface of the lung is considerably reduced, as in pneumothorax, atelectasis, pulmonary embolism, etc. The importance of the reflex can be demonstrated by cutting both vagi, an operation which immediately suppresses rapid shallow breathing. Hypersensitiveness of the pulmonary stretch receptors seems to be the cause of this type of dyspnea.

Cardiac asthma and orthopnea. Cardiac dyspnea is observed in certain cases of heart disease in which there is congestion (and sometimes edema) of the lung. The decrease in vital capacity, and probably in pulmonary distensibility, diminishes the amplitude of the respiratory movements. The Hering-Breuer reflex is exaggerated, and shallow breathing causes insufficient oxygenation of the blood. Dyspnea is the result of all these factors. In certain forms of severe cardiac insufficiency, especially if there is hypertension, paroxysmal attacks of dyspnea occur, usually during the night; this is known as cardiac asthma. Morphine diminishes the severity of the attacks of cardiac asthma, because this drug depresses the respiratory center.

Dyspnea increases when the patient lies down and diminishes when the thorax is in a vertical position. This is known as orthopnea. The improvement in breathing observed in this position has been interpreted as due to the following mechanism: when the thorax is in the orthopneic position (*i.e.*, vertical) the abdominal viscera do not press against the diaphragm, and the negative pleural pressure increases; this causes pulmonary congestion to diminish. The distensibility of the lung and vital capacity increases; this causes the Hering-Breuer reflex to diminish; circulation improves, the pressure of the cerebrospinal fluid diminishes, and the blood supply to the respiratory center also improves.

APNEA

Apnea literally means absence of breathing, but the term is used to signify temporary suspension of breathing. Apnea can be provoked by several causes, the most common being acapnia, *i.e.*, a fall in the CO_2 concentration of the plasma due to hyperventilation (Fig. 153).

Fig. 153. Pneumogram of a dog. Chloralose anesthesia. The arrow marks intravenous injection of 0.1 mg. nicotine per kg. There is hyperpnea followed by apnea. Time in minutes.

Hyperventilation, obtained by forced breathing during 2 min., provokes a series of responses due to the alkalinization of the blood. One of these is that the need to breathe is not felt for several minutes while CO_2 accumulates and again reaches a certain concentration. This fact is used by divers to prolong the time they can remain under water. If hyperventilation is carried out in an atmosphere of pure oxygen, apnea can be prolonged for an even longer time.

Apnea due to acapnia can be provoked experimentally in animals, and it has been amply demonstrated that it is due to a decrease in CO_2 concentration and not to a decrease in the excitability of the respiratory center or to a reflex. Hyperventilation with air or oxygen containing an adequate CO_2 concentration is not followed by apnea. Section of both vagi, so as to suppress the Hering-Breuer reflex, does not prevent apnea after hyperventilation. The importance of CO_2 concentration has also been demonstrated in crossed-circulation experiments. The head of a dog *A* is supplied with blood from the body of a second dog *B*, and the body of *A* supplies the head of *B*. Hyperventilation of *B* causes apnea in *A*; and the accumulation of CO_2 (hypercapnia) in *A*, due to the stopping of breathing, causes hyperventilation in *B*. Perfusion of an isolated head gives even more spectacular results, because the only blood supply to the perfused head comes from the perfusing animal, while in the crossed-circulation experiments some blood reaches the head from the body of the same animal by way of the arteries in the spinal canal. Perfusion of the head with hypercapnic blood causes ample and rapid movements of the alae nasi, and if asphyxia is pronounced, gasping movements of the jaw. Perfusion with acapnic blood causes the movements of the alae nasi to cease.

Apnea due to hyperventilation is of practical importance in general anesthesia, especially in ether anesthesia. During the initial phase there is hyperexcitability of the respiratory center and hyperventilation occurs. Later, as anesthesia progresses, the excitability of the respiratory center diminishes and returns to normal or below. Acapnia provoked by the initial hyperventilation is more marked as the sensitiveness of the respiratory center is depressed. Anoxia can thus be provoked, and precautions should be taken to avoid it. This can be done by adding 5 to 8 per cent CO_2 and oxygen to the anesthetic gases. Usually the patient is made to rebreathe his own expired air, which contains CO_2.

During apnea due to acapnia there are other symptoms of alkalosis: dizziness, sometimes vomiting, and signs of neuromuscular hyperexcitability, such as exaggeration of reflexes and

spontaneous muscular contractions, *i.e.*, tetany due to alkalosis.

Apnea is also caused by a decrease in the excitability of the respiratory center. Thus intravenous injection of a sufficient dose of morphine into a dog produces apnea, which lasts until CO_2 accumulates and reaches a high enough concentration to stimulate the depressed respiratory center.

Sudden arterial hypertension causes transitory apnea due to a reflex initiated in the vasosensory areas.

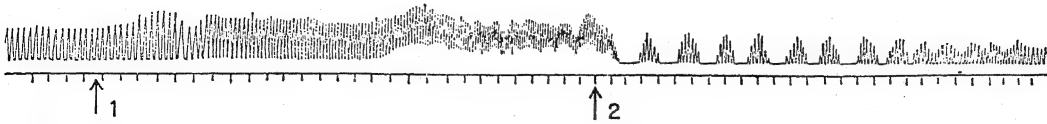



FIG. 154. Periodic breathing. Pneumogram of subject with Cheyne-Stokes breathing due to hyperpnea caused by anoxia. 1, a gas mixture containing only 9.84 per cent oxygen is breathed, evoking hyperpnea; 2, the subject breathes air. Time in 10 sec. (*Haldane, J. S., and J. G. Priestley, "Respiration," Oxford, New York, 1935, p. 196.*)

PERIODIC BREATHING

Periodic breathing is an involuntary disturbance in respiratory rhythm. It can be of several types, but the most common is known as Cheyne-Stokes breathing. This consists in the regular alternation of periods of breathing separated by periods of apnea. Breathing commences by shallow movements, which gradually increase in amplitude and sometimes become dyspneic. Respiratory movements then gradually decrease in amplitude, and eventually cease. After a period of apnea another cycle commences. The duration of the periods of apnea and breathing varies considerably in different cases (Fig. 154).

There are other types of periodic breathing. Thus in Biot's breathing, the periods of breathing begin and end suddenly, without the gradual increase and decrease in the amplitude of the respiratory movements typical of Cheyne-Stokes breathing; there is also considerable irregularity in the duration of the different periods. Biot's breathing is observed in meningitis and has a more serious significance than Cheyne-Stokes breathing.

The periods in Cheyne-Stokes breathing are nearly always accompanied by periodic variations in arterial blood pressure, which falls during apnea, and rises during the period of breathing. Sometimes patients with Cheyne-Stokes breathing lose consciousness during the apneic periods.

Cheyne-Stokes breathing occurs in many different circumstances. It is sometimes seen in normal infants, during sleep, especially in premature babies; it is rarely observed in normal adults. It is a frequent occurrence in the course of adaptation to high altitudes, especially during sleep. It is also observed after apnea caused by hyperpnea. It is normal in aquatic mammals (seals, etc.),[1] and in hibernating animals during the winter sleep.

Cheyne-Stokes breathing may be a sign of anoxia due to cardiac decompensation or of pulmonary origin (pneumonia, etc.), of toxic depression of the respiratory center (morphine, anesthesia), of uremia, or of lesions of the nerve centers (endocranial hypertension).

The fundamental factor in periodic breathing is anoxia of the medullary centers; probably there is also depression of the respiratory and other brain centers. There are still many unexplained points in the mechanism of periodic breathing. The following interpretation of Cheyne-Stokes breathing has been widely accepted: Anoxia diminishes the excitability of the respiratory center and provokes apnea. This increases anoxia sufficiently to stimulate the chemoreceptors in the carotid and aortic bodies, and the excitability of the center is increased reflexly; during the apneic period CO_2 is accumulated and reaches a concentration high enough to stimulate the respiratory center. Respiration is reestablished; therefore anoxia and CO_2 concentration diminish, and again apnea occurs. The breathing of oxygen or air with CO_2 usually suppresses periodic breathing.

Other centers besides the respiratory center play a part in the establishment of periodic breathing. This is evident by the greater ease with which this type of breathing takes place during sleep or unconsciousness, and under the effects of drugs that depress the central nervous

[1] SWINDLE, P. F., *Am. J. Physiol.*, **79**, 188, 1926.

system. Periodic breathing is difficult to provoke in the dog, but it is more easily established after decerebration. Swindle maintains that there are spinal centers that emit periodic respiratory impulses, normally controlled by the respiratory center in the medulla. In the newborn the medulla has not yet acquired complete control over the spinal centers; hence the frequent occurrence of Cheyne-Stokes breathing in infants. The importance of other nerve centers and the mechanism by which they take part in the production of periodic breathing are not yet well understood.

ANOXIA

When there is not a sufficient amount of oxygen to satisfy the normal requirements of cells, a state of anoxia is established.

Anoxemia literally means lack of oxygen in the blood. Anoxia has a wider significance, because in some cases oxygen insufficiency in the tissues, or the disturbance in the utilization of oxygen, is not due to lack of oxygen in the blood. "Hypoxia" would be a more correct term, as vertebrates cannot survive total lack of oxygen for more than a very few minutes; but "anoxia" is currently used and it should be understood to signify not complete absence of oxygen but oxygen insufficiency. The use of a word with a meaning different from its original or literal meaning often becomes widely accepted; *e.g.*, "anemia," which literally means absence of blood, and "oxygen," which means a generator of acids, are words commonly used with a somewhat different significance. When a term is thus in current use, it should be accepted, provided that the new meaning is well defined.

Anoxia is frequently accompanied by changes in CO_2 concentration and alkali reserve. Special terms are used in these cases. "Acapnia" (Greek α, negative prefix, and *καπνος*, smoke) signifies a decrease in the CO_2 concentration of the blood. "Hypocapnia" has the same meaning; "hypercapnia" signifies an increase in blood CO_2 concentration. "Acarbia" ("hypocarbia") and "hypercarbia" mean respectively decrease and increase in alkali reserve. "Asphyxia" literally means "not to throb" (Greek α, negative prefix, and *σφυξειν*, to throb). It is frequently used incorrectly as a synonym for "anoxia"; it means lack of oxygen plus retention of CO_2, *i.e.*, anoxia plus hypercapnia.

Classification of anoxia. There are four types of anoxia, according to the origin of the disturbance: (*a*) anoxic anoxia, due to insufficient oxygenation of the blood in the lungs; (*b*) anemic anoxia, due to a lowered oxygen capacity of the blood; (*c*) stagnant anoxia, due to circulatory disturbances that prevent the rapid circulation of oxygenated blood in the capillaries; (*d*) histotoxic anoxia, due to poisoning of the oxidizing mechanisms of the tissues. These four types will be considered first; then a more detailed study will be made of certain practically important forms of anoxia.

Anoxic anoxia. Venous blood normally retains a considerable amount of oxygen (see Chap. 27). In a normal subject at rest arterial blood loses only 6 volumes per cent oxygen on passing through the capillaries; venous blood has 14 volumes per cent oxygen. The important factor in this exchange of O_2 between HbO_2 and the tissues is not the volume of gas but the difference in the partial pressures. The oxygen tension in arterial blood is 100 mm. Hg; in venous blood it is 37 mm. Hg. Venous blood has a volume of oxygen amply sufficient for all the needs of a resting organism, but it is at a low partial pressure so it is not easily given up to the tissues.

Oxygen tension is not the only factor conditioning its passage from the lung to the blood and from the blood to the tissues; an increase in CO_2 concentration favors the release of oxygen from hemoglobin. This factor has special importance in anoxia at high altitudes.

At the normal oxygen tension of arterial blood (100 mm. Hg) 95 per cent of the hemoglobin is in the state of HbO_2. When the O_2 tension and the percentage of HbO_2 are below these figures, there will be anoxic anoxia.

Anoxic anoxia is due to the following causes:

1. *Low oxygen tension in inspired air.*
   - *a.* Decrease in total pressure such as occurs at high altitudes, in decompression chambers for reproducing conditions in high altitudes, etc.
   - *b.* Respiration of inert gases.
   - *c.* Breathing in a confined space, such as that of submarines, mines, etc., in which oxygen is consumed; increase in the dead space; breathing in a closed circuit, as in administration of anesthesia with ethylene or nitrous oxide, which must be breathed in high concentration.

2. *Disturbances in pulmonary exchange of gases; hypoventilation.*

   a. Obstruction of the air passages (larynx, trachea, bronchi), *e.g.*, drowning, asthma.
   b. Pulmonary disturbances that prevent alveolar ventilation: alveolitis, edema, toxic gases, emphysema, pneumonia, atelectasis, etc.
   c. Inadequate expansion of the lung, of pulmonary or extrapulmonary origin: pneumothorax, pleural effusion, emphysema, pulmonary fibrosis, atrophy of the respiratory muscles, etc.
   d. Insufficient pulmonary ventilation, due to depression of the respiratory center.
   e. Shallow breathing due to shock, pneumonia, bronchopneumonia, postoperative pain, neurasthenia, fatigue, etc.

3. *Short circuit between the right and left sides of the heart or in the pulmonary circulation.* In such cases the blood passing through the lung is well oxygenated but is mixed with venous blood that has not passed through the lung:

   a. Congenital heart disease, interauricular or interventricular septal defects, etc.
   b. Arteriovenous aneurysms in the pulmonary circulation.
   c. Passage of blood through parts of the lung that are not ventilated, pneumonia, atelectasis.

4. *Retarded diffusion of gases due to alterations in the alveolar epithelium* and perhaps in the vascular endothelium.

Anemic anoxia. In this type of anoxia the gaseous exchanges in the lung are carried out normally, and the oxygen tension in the blood is normal, but the oxygen capacity of the blood is below normal because there is a decrease in hemoglobin or because the latter is altered or is partially combined with some gas other than oxygen; the total volume of oxygen is therefore diminished. If the condition is not very severe, the amount of oxygen in the blood can be sufficient in resting conditions, but oxygen want is made evident in muscular exercise because the oxygen reserve is below normal. This type of anoxia is due to the following causes:

1. *Hemoglobin insufficiency,* acute and chronic hemorrhage, severe anemia.
2. *Alteration of hemoglobin,* formation of methemoglobin in cases of intoxication by nitrite, chlorate, aniline, etc.
3. *Combination of hemoglobin with gases other than oxygen,* CO intoxication (see page 30).

Stagnant anoxia. This type of anoxia results from slow circulation of the blood in the capillaries due to general or local causes. Blood is normally oxygenated in the lungs, but it circulates slowly through the capillaries and the tissues do not receive adequate amounts of oxygen. Owing to the slow circulation, the CO_2 content of the blood increases abnormally, the release of O_2 from hemoglobin is thus facilitated, and the want of oxygen is partially compensated. Stagnant anoxia can affect the whole organism, or be localized to one vascular territory. It is due to the following causes:

1. *Central circulatory disturbance,* heart insufficiency.
2. *Peripheral circulatory disturbance,* circulatory collapse.
3. *Local anoxia,* local stenosis or arterial spasms, venous obstruction.

Histotoxic anoxia. In this type of anoxia the oxygenation and circulation of the blood are normal but the oxidative mechanisms of the tissues are disturbed. Any drug that depresses cell respiration causes it; the most typical is cyanide, but H_2S, narcotics, and other drugs also produce it. The particular ways in which the different drugs act on the oxidation-reduction mechanism are described in treatises on cell respiration.

Symptoms of anoxia. Anoxia can be either acute or chronic. The signs and symptoms of acute anoxia are similar to those of alcohol intoxication; those of chronic anoxia resemble those of fatigue.

Mammals, including man, do not tolerate acute anoxia; after a few minutes in an atmosphere free from oxygen or with a low oxygen tension, severe symptoms become evident, and death follows. If the subject recovers, permanent injury may remain. Anoxia, therefore, not only stops the organic functions but also damages them. Organs most sensitive to anoxia are the central nervous system and the heart.

Nervous symptoms: headache, depression, apathy, slowness of thought, or else excitement, loss of control, loss of memory and the idea of time,

intellectual errors, emotional reactions, inefficient muscular coordination, asthenia, weakness, fatigue, somnolence or insomnia. In an advanced stage there are disturbances in the senses (vision, hearing, pain, etc.).

Digestive symptoms: nausea, vomiting.

Circulatory symptoms: tachycardia, slight hypertension, cyanosis, disturbances in the electrocardiogram, nasal hemorrhage. In advanced stages: heart failure, dilatation of the heart, bradycardia, sudden fall in blood pressure, syncope.

Respiratory symptoms: increased pulmonary ventilation, tendency to periodic respiration, especially during sleep.

Permanent sequelae of severe anoxia are due to degenerative lesions in the central nervous system, especially in the cortex and the subcortical nuclei.

Special Forms of Anoxia

Anoxia can be established rapidly (acute anoxia) or gradually (chronic anoxia). In chronic anoxia there are compensatory reactions that adapt the organism to the lower oxygen tension. These have been well studied in the acclimatization to high altitudes.

Mountain sickness. This disturbance is due to anoxic anoxia; it is called "*soroche*" in Peru and "*puna*" in the plateaus of the southern Andes.[1] As altitude increases the barometric pressure falls (see Chap. 27); therefore oxygen tension also falls, and the percentage of oxyhemoglobin in blood diminishes. The signs and symptoms of mountain sickness are due to the decrease in oxygen tension. Exercise increases the severity of the condition because it increases the need for oxygen.

A rapid ascent to high altitude in an airplane provokes symptoms similar to those produced by sudden decompression in divers. In both cases gaseous nitrogen is released in the tissues (see "Accidents due to decompression," page 313). In the pathogenesis of these disturbances a rapid fall from 4 atmospheres to 1 atmosphere has the same effect as a drop from 1 to ¼ atmosphere.

Mountain sickness appears at different oxygen tensions according to whether the subject has become acclimatized or not, or is active or resting; there are also considerable individual variations. The first signs appear at heights between 9,000 and 10,000 ft. (3,000 m.). Oxygen tension at 10,000 ft. is 110 mm. Hg and it is 66 mm. Hg at 22,000 ft. Higher altitudes can be reached by unacclimatized persons without danger only if oxygen is breathed; usually aviators begin to breathe oxygen at even lower levels.

In 1875 three French scientists ascended in a balloon to a height of 26,000 ft. (8,600 m.); at about 25,000 ft. (7,500 m.) consciousness was lost and only one of them, Tissandier, survived. After several unsuccessful attempts, in some cases with loss of lives, on May 29, 1953, the peak of Mt. Everest, slightly over 29,000 ft., and the highest on earth, was reached by Hillary and Tenzing, members of the British Mt. Everest Expedition, led by Sir John Hunt. The last thousand feet were the most difficult to climb, and it was necessary to breath pure oxygen (3 to 4 liters per min. in an open circuit) in order to keep going.[1]

The response of the organism to high altitude has been the subject of considerable research in permanent and temporary laboratories established at great heights in different parts of the world. Mosso organized the laboratory on the Gnifetti peak of the Monte Rosa in the Italian Alps at 15,000 ft. The Peak of Tenerife (11,000 ft.), Pike's Peak (14,000 ft.), and the Cerro de Pasco in Peru (14,000 ft.) have been the sites of famous expeditions. Work in the laboratories on the Jungfrau and Davos in Switzerland, and at the camp at Auncanquilcha (20,000 ft.) in Chile has also given many interesting results. The University of San Marcos, Lima, Peru, has an Institute of Andean Biology with a permanent laboratory situated 90 miles from Lima, at Morococha, which is 14,900 feet above sea level. The University of California operates the White Mountain high-altitude research station, with laboratories at 10,600 and 12,500 ft. The effect of low pressure has also been studied in hermetically sealed steel chambers in which a pneumatic pump can produce any required pressure (decompression chambers). These decompression chambers are in use at the principal aviation fields to test the candidates for aviation pilots. They cannot be used for studies of acclimatization, as this requires days or even weeks.

[1] Mountain sickness is observed in subjects living at 10,000 ft. (3,000 m.) or more above sea level. Aviation sickness is due to sudden and repeated exposures to low barometric pressures for a short time.

[1] Hunt, J., "The Ascent of Everest," Hodder and Stoughton, London, 1953.

There is wide variation in the responses of different individuals. They begin at lower levels in some subjects than in others. Physical exercise, such as is performed when climbing a mountain, exposure to cold, wind, etc., increase the need of oxygen and accelerate the appearance of in oxygen tension and not the decrease in barometric pressure.

Other signs and symptoms that appear at about 10,000 ft. are tachycardia, sometimes premature beats, increased respiratory frequency, fatigue, insomnia, etc. After 2 to 4 days'

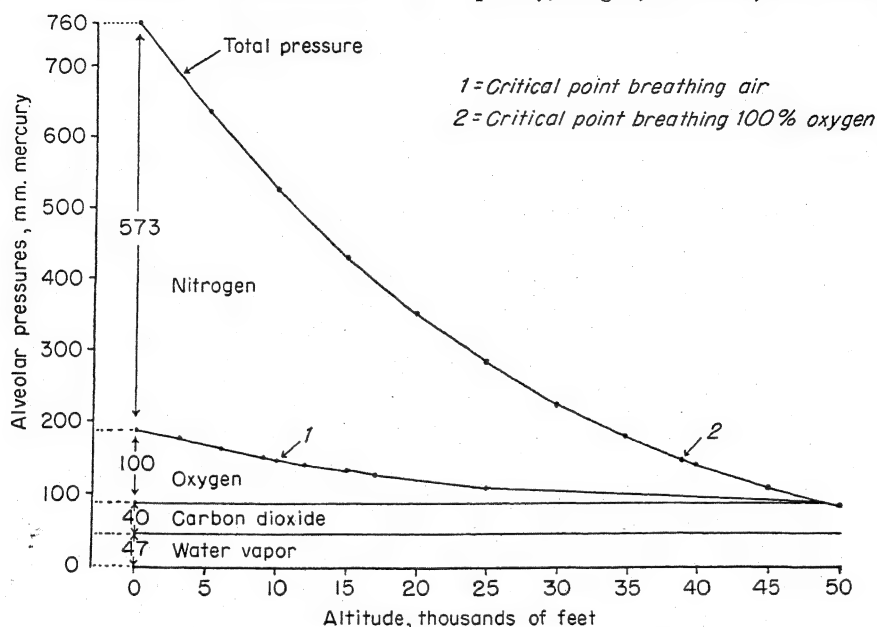



FIG. 155. Total pressure and partial pressure of gases in the alveolar air at different altitudes. (*Fulton, J. S., et al. "Decompression Sickness," Saunders, Philadelphia, 1951.*)

the response. If pure oxygen is breathed instead of air, the critical point of oxygen insufficiency is shifted upward; instead of 9,000 to 10,000 ft. (3,000 m.), it is 38,000 ft. (12,000 m.). CO_2 partial pressure and tension of water vapor remain constant (Fig. 155).

The disorders arising in the nerve centers are due to cerebral anoxia. There is loss of memory, errors are committed when making simple arithmetical calculations, and there are also errors of judgment. For instance, when Haldane wished to observe his own cyanosis while being submitted to a total pressure of 320 mm. Hg in a decompression chamber, he kept searching for his image on the back of the mirror. Dizziness, headache, and vomiting then occur, and as the barometric pressure falls the subject becomes excited. If the pressure is lowered even more, progressive paralysis makes its appearance and death may occur.

These disorders are not observed if the subject breathes pure oxygen instead of air at these same pressures; they are therefore due to the decrease

exposure to this altitude the symptoms begin to recede gradually.

The decrease in the oxygenation of the blood causes cyanosis, which is easily observable in the lips and nails and rapidly increases with exercise.

One of the first symptoms to appear is an increase in pulmonary ventilation, due at first to increase in depth, and later also in frequency, of breathing. This is observed when the oxygen falls to the equivalent of 13 per cent at sea level, but in some cases it begins at an oxygen partial pressure corresponding to 18 per cent oxygen at sea level. This increase in pulmonary ventilation is caused by an increase in the excitability of the respiratory center due to the stimulation by anoxia of the chemoreceptors in the aortic and carotid bodies. It increases the elimination of CO_2, and CO_2 concentration falls in the alveolar air and in the blood (acapnia). The loss of free CO_2 modifies the ratio of free CO_2 to combined CO_2, therefore the hydrogen ion concentration in the blood diminishes. This disturbance in the acid-base equilibrium is com-

pensated by the elimination of base by the kidneys, until the normal ratio is restored. Acclimatized individuals have a lower alkali reserve than normal at sea level. This does not mean that the acid-base equilibrium is disturbed in high altitudes, but that it is adjusted to the new conditions in which the organism finds itself. This decrease in alkali reserve is not, therefore, a condition of acidosis. A subject acclimatized to an altitude of 15,000 ft. has different physiologic constants from those normal for sea level.

Acclimatization to high altitude, or conversely from high altitude to sea level, takes place gradually. With respect to the acid-base equilibrium and respiratory conditions, the adjustment takes about a month to establish.

Normal CO_2 tension in alveolar air is more or less 40 mm. Hg at sea level; at an altitude of 14,000 ft. it falls gradually and becomes stabilized at 26 mm. Hg after 35 days. The respiratory minute volume at rest remains stable at a figure 50 per cent above that at sea level. In acclimatized subjects there is a close relationship between the level to which they are adapted and the alveolar CO_2 tension.

Considering only these factors, the unacclimatized individual is at a disadvantage for three reasons: (*a*) he has a lower oxygen concentration in arterial blood; (*b*) he has a lower oxygen tension; (*c*) his blood is more alkaline, a condition that is unfavorable for the dissociation of oxygen from oxyhemoglobin.

The decrease in alveolar CO_2 tension permits a corresponding increase in alveolar oxygen tension, which compensates in part the anoxic condition.

Individuals of a community that has lived at high altitudes for several generations show another type of adaptation. Barcroft reports that the Cholos who live on the Peruvian plateau have ribs placed more horizontally, so that the transverse diameters of the thorax are relatively larger and the vertical diameter less than in persons living at sea level; the thorax thus has a larger volume in the inhabitants of the plateau.

Another important change in acclimatization is the increase in erythrocyte (polycythemia) and hemoglobin concentration. Polycythemia is observed immediately on exposure to high altitude; this is due to contraction of the spleen caused by anoxia, a fact that has been demonstrated experimentally. The total amount of red cells in the organism is not increased, but those stored in the spleen are poured out into the circulation, and there is an increase in the circulating blood volume. Polycythemia persists in subjects living at high altitude, and many theories have been put forward to explain this so-called secondary polycythemia, such as a longer life span of the red cells, an increase in blood concentration due to the evaporation of water, etc. The true cause of polycythemia is an increase in the activity of the hemopoietic tissue in the bone marrow. Chronic anoxia of other origins, such as chronic CO intoxication, also provokes polycythemia. This greater activity of the bone marrow is made evident by an increase in reticulocytes. Regeneration of red cells after experimental hemorrhage in dogs also takes place at a faster rate in animals at high altitudes than in those at sea level. After a few weeks, sometimes months, of living at high altitude the erythrocyte count is stabilized at a level above the normal for sea level; there is a certain relationship between the erythrocyte and hemoglobin concentration, on the one hand, and the height above sea level, on the other.

Barcroft found an increase in the oxygen-carrying capacity of hemoglobin in persons living at high altitudes, but this fact has not been confirmed by other workers. Hurtado has observed an increase in myohemoglobin in acclimatized persons.

Monge[1] has described a chronic form of mountain sickness in which the patients cannot adapt themselves or have lost the capacity of adaptation to high altitude. The abnormal condition persists as long as the subjects remain at high altitude, and rapidly disappears on their return to sea level. Irreversible damage (heart disease, etc.) may be suffered if such persons live several years at high altitude without achieving satisfactory acclimatization.

Decrease in oxygen concentration. This condition is found at normal or even high barometric pressure, in a closed environment such as mines, if the air is vitiated by inert gases, such as methane. Signs and symptoms of anoxia appear when the oxygen concentration falls to 13 per cent. In mines, CO is also formed, but this causes anoxia by a different mechanism.

Anoxia due to pulmonary disease. Any disease of the lung that diminishes the surface of gaseous exchange by a certain extent causes

[1] MONGE, D., *Science*, **95**, 79, 1942.

anoxia. Pulmonary diseases are particularly serious where the barometric pressure is low, because two factors that cause anoxia are present at the same time. Pneumonia at an altitude of 15,000 or 20,000 ft. is almost certainly fatal, and if possible the patient should be brought down to sea level, or at least adequately treated with oxygen.

Pulmonary diseases cause anoxia in several ways. In pulmonary edema the alveoli and small bronchi are obstructed by fluid; in bronchopneumonia they are filled with secretions. In asthma air circulates with difficulty because of spastic contraction of the smooth muscles of the small bronchi. In pneumonia the consolidated area does not receive air; during the first stages the blood vessels of the diseased area are still patent but the blood does not come in contact with the air, so it returns to the heart in a venous condition, where it is mixed with blood which has been arterialized in the normal areas. There is also shallow breathing in these patients, and therefore insufficient alveolar ventilation. In certain chronic diseases of the lung, such as emphysema, the alveolar walls suffer structural changes which obstruct gaseous interchange. In other chronic diseases, such as Ayerza's disease, chronic anoxia provokes some of the effects observed in mountain sickness, *e.g.*, polycythemia due to stimulation of the hemopoietic tissue in the bone marrow.

Fetal anoxia. According to Barcroft, who has extensively studied fetal respiration, fetal blood is normally in a state of anoxia, due to the conditions of gaseous interchange in the placenta. The fetus has polycythemia provoked by this anoxia, but after birth rapidly becomes "acclimatized" to the new condition of life.

CYANOSIS

Cyanosis is a dark blue or violet color of the skin and mucosae, observed when the blood in the surface capillaries contains a certain proportion of nonoxygenated hemoglobin, or hemoglobin compounds such as methemoglobin or sulfhemoglobin. Not all compounds of hemoglobin produce cyanosis; thus in CO intoxication the skin is pink, because of the cherry-red color of HbCO.

Cyanosis is observed when 5 gm. per cent or more of the hemoglobin in the capillaries is in the nonoxygenated state. Sometimes cyanosis is provoked by 4 gm. per cent of hemoglobin in this condition and other times 6 gm. per cent is needed.

There are four factors in the production of cyanosis: (*a*) total hemoglobin concentration in blood; (*b*) degree of oxygenation of the blood leaving the lungs; (*c*) the proportion of arterial to venous blood when these are mixed, as in congenital cardiac malformation; (*d*) removal of oxygen from the blood as it passes through the capillaries.

The intensity of cyanosis is modified by certain factors that facilitate or obstruct the observation of the color of the blood in the capillaries: thickness of the skin overlying the capillary network; diameter, length, and number of capillaries in a given area of skin or mucosa; pigmentation of the skin, normal (Negroes, Mongols) or abnormal (deposits of silver or gold, Addison's disease, jaundice); conditions of the blood plasma (pigments, chylomicrons, leukocytosis, etc.). Cyanosis can be more easily detected in areas where the skin is fine, poorly pigmented, and well vascularized, *e.g.*, lips, nose, ears, cheeks, finger tips, and especially over the root of the nails. In the course of surgical operations the color of the blood can be noted in the viscera or directly as it flows from the blood vessels.

The intensity of cyanosis is due essentially to the absolute amount of nonoxygenated hemoglobin; therefore the best way of measuring it is by determining the amount of nonoxygenated hemoglobin in capillary blood. Total hemoglobin is determined, then the percentage of nonoxygenated hemoglobin in arterial and in venous blood; the average of these figures is the one corresponding to capillary blood.

For example, in a subject with normal blood, the total oxygen capacity is 20 cc. per cent. Arterial blood carries 19 cc. per cent O_2; therefore there is the equivalent of 1 cc. per cent nonoxygenated hemoglobin. As 1 gm. hemoglobin carries 1.34 cc. O_2, there is 0.74 gm. hemoglobin. Venous blood carries 14 gm. per cent oxygen; there is therefore the equivalent of 6 cc. per cent (4.47 gm.) nonoxygenated hemoglobin. In the capillaries there is $(1 + 6)/2 =$ 3.5 cc., *i.e.*, 2.6 gm. hemoglobin. This figure is below the threshold of cyanosis.

In a subject with cyanosis the total oxygen capacity is 20 cc. per cent. Arterial blood carries 16 cc. per cent; the equivalent of 4 cc. per cent remains unsaturated. Venous blood carries 10 cc. per cent; the

equivalent of 10 cc. per cent remains unsaturated. In capillary blood there is $(4 + 10)/2 = 7$ cc., which corresponds to 5.22 gm. nonoxygenated hemoglobin.

Cyanosis is closely connected with anoxia, but the two terms do not have the same significance. For example, in severe anemia there is anoxia, but no cyanosis, because there is not enough hemoglobin. In polycythemia there can be cyanosis without anoxia. In anoxia due to CO intoxication and in histotoxic anoxia there is no cyanosis. "Anoxia" and "cyanosis" are not, therefore, equivalent terms.

Classification of cyanosis. *Arterial* or *central cyanosis.* In this type the blood passes through a normally ventilated lung but is not sufficiently oxygenated. The condition can be provoked by

1. Decrease in oxygen tension in inspired air, due to
   - *a.* Breathing air contaminated with inert gases.
   - *b.* Breathing in closed atmospheres where oxygen is consumed and inadequately replaced (mines, submarines).
   - *c.* Low barometric pressure (high altitudes; decompression chambers).
2. Insufficient alveolar ventilation, due to
   - *a.* Shallow breathing (thoracic pain, pneumonia, bronchopneumonia, encephalitis, hysteria, neurasthenia, postoperative conditions, etc.).
   - *b.* Diminished respiratory rate (narcosis, encephalitis, etc.).
   - *c.* Obstruction of the air passages (larynx, trachea, bronchi), asthma, bronchopneumonia, etc.
   - *d.* Diminished pulmonary expansion (narcosis, emphysema, sclerosis, Ayerza's disease, pleural effusion, pneumothorax, etc.).
3. Short circuit between venous and arterial circulation, due to
   - *a.* Short circuit between the right and left sides of the heart (interauricular or interventricular orifices; pulmonary arteriovenous fistula, etc.).
   - *b.* Passage of blood through unventilated areas of the lung that have patent blood vessels (atelectasis, pneumonia, pulmonary congestion, edema, etc.).

In all the cases listed under 1 and 2, the administration of oxygen diminishes cyanosis. In the cases listed under 3, the administration of oxygen does not diminish cyanosis except in cases of pneumonia in which cyanosis is due mainly to shallow breathing (2*a*).

Peripheral cyanosis. In this type the blood loses a large proportion of oxygen to the tissues, owing to slow circulation through the capillaries or to anoxic anoxia. If there is circulatory stasis, there is also venous hypertension. Peripheral cyanosis can be general or local. It is due to

1. Congestive cardiac decompensation. If pulmonary congestion is marked, central cyanosis, due to insufficient oxygenation of the blood, will be added.
2. Venous obstruction (mediastinal tumor, mitral stenosis, stenosis of the pulmonary artery, etc.).
3. Peripheral circulatory insufficiency (circulatory collapse in shock, infections, intoxications, etc.).
4. Arteriolar or venous spasmodic vasoconstriction, with capillary dilatation (Raynaud's disease, cold, acrocyanosis, etc.).
5. Combination of two causes of anoxia, or of anoxia with increased oxygen consumption (physical exercise at high altitudes; cardiac and pulmonary diseases at high altitudes, which are well compensated at sea level).

Cyanosis of exclusively peripheral origin is not improved by oxygen. Local cyanosis due to vascular spasm is improved by submerging the affected limb in water at 45°C., which relieves the spasm.

Cyanosis due to alterations in hemoglobin. This type of cyanosis is due to

1. Formation of methemoglobin (intoxication by sulfanilamide, aniline, nitrobenzene, nitrites, etc.).
2. Formation of sulfhemoglobin (intoxication by H_2S, enterogenous autotoxic cyanosis or Stovkis-Talma syndrome).

THERAPEUTIC USE OF OXYGEN AND CARBON DIOXIDE

Pure oxygen at a pressure of 4 to 5 atmospheres is toxic and rapidly fatal. The majority of animals die after a few days in an atmosphere with 70 to 80 per cent oxygen, but 50 per cent is well tolerated. Breathing more than 60 per cent oxygen at normal pressure is toxic for man after a few hours. The onset and severity of

symptoms depend on the oxygen partial pressure in the air inspired, the duration of exposure, and individual sensitiveness; this last factor has considerable importance in modifying the severity of the symptoms. Pure oxygen should, therefore, not be given for more than a short time. Moreover, it is usually not necessary to do so, because satisfactory results are obtained with 60 per cent oxygen in all but a few exceptional cases.

Breathing pure oxygen at sea-level pressure for several hours produces irritation of the respiratory tract in normal subjects. There is also retrosternal pain, and the vital capacity diminishes. In 82 per cent of the subjects retrosternal pain appears in 4 to 22 hr. of exposure to pure oxygen.[1] This symptom gradually subsides when air is again breathed; usually it disappears in less than 12 hr. Breathing 75 per cent oxygen provokes symptoms in 55 per cent of the subjects, but none is provoked when 50 per cent oxygen is breathed. The reactions are due exclusively to the high oxygen tension. Breathing air at high pressure does not produce them. They are not due to lack of nitrogen, because breathing pure oxygen in a decompression chamber at 380 mm. Hg does not provoke them. At this pressure pure oxygen is at the same partial pressure as 50 per cent oxygen at sea-level pressure.

Breathing pure oxygen at 2 or more atmospheres causes predominantly signs of disturbance of the cerebral cortex, *e.g.*, epileptoid convulsions. In experimental animals, permanent lesions and death have been observed.

The administration of oxygen at concentrations higher than in air increases the oxygen tension in alveolar air and its diffusion into the blood; hence its efficiency in the treatment of certain types of anoxia.

Breathing oxygen is highly beneficial when the oxygen tension is low, as in high altitudes (aviation), when the air is vitiated (mines), when there is mechanical obstruction of the air passages (drowning), in respiratory depression (circulatory collapse, anesthesia), when gaseous exchange in the lung is obstructed (pneumonia), in pulmonary edema (poison gases, heart disease), in carbon monoxide intoxication, whenever artificial respiration must be performed, in emphysema, and in asthma. In all these cases the higher oxygen tension increases oxygen diffusion into the blood and a higher degree of oxygenation of hemoglobin is assured.

Breathing oxygen has no beneficial effects in other forms of anoxia. In anemia, and in cases of intoxication with conversion of hemoglobin into methemoglobin, anoxia is due to the decrease of hemoglobin, but the hemoglobin present is 95 per cent saturated with O_2; oxygen inhaled does not significantly increase this proportion nor the amount of O_2 dissolved in the plasma. Oxygen therapy is not efficacious in stagnant anoxia unless there is extensive congestion or edema of the lungs. In stagnant anoxia, hemoglobin leaves the lungs normally oxygenated, so any increase in the alveolar oxygen tension will have little effect. In histotoxic anoxia (*e.g.*, cyanide intoxication), there is no oxygen want, but there is incapacity to utilize oxygen; therefore in these cases oxygen therapy is not efficacious.

Oxygen therapy should not be applied in chronic anoxia (except in the acute stages) because it interferes with the activity of compensatory mechanisms, and after oxygen administration is discontinued the patient is in worse condition than before the treatment.

In cases in which oxygen therapy is efficacious the patient improves rapidly, dyspnea ceases, and the circulation improves. If there is fever, as in pneumonia, the temperature falls.

Oxygen therapy is not efficacious if the oxygen is not given in the proper conditions.[1] It should be administered continuously over several hours, and sometimes days. The oxygen reserve in the organism is sufficient for only a few minutes; and this reserve is diminished in anoxia. Suppression of oxygen, even for a few minutes, can be more harmful than the complete absence of oxygen therapy, because in the latter condition compensatory mechanisms are active, which oxygen therapy suppresses.

Oxygen should be given in an adequate concentration. The use of a funnel or a tube placed near the nasal orifices is not efficacious. Oxygen masks are useful in certain cases such as asphyxia due to CO poisoning, drowning, anoxia of aviators flying at high altitudes, and when the treatment is necessary for only a few minutes or, at most, a few hours. In the treatment of anoxia due to disease, when oxygen must be administered over long periods, oxygen tents (Fig. 156)

[1] Comroe, *et al.*, Oxygen Toxicity, *J. A. M. A.*, **128**, 710, 1945. (A very good bibliography of the subject is given.)

[1] It is not necessary to administer the 99 to 99.5 per cent "medicinal" oxygen; 98 per cent oxygen prepared for industrial uses is quite as efficient, has no drawbacks, and is considerably less expensive.

or masks that cover all the face must be used, or a catheter should be placed inside the nasal cavity.

Oxygen tents are made of impermeable material with plastic windows. They are hung over the head of the patient's bed and the folds tucked beneath the mattress. The patient consumes only 0.5 liters per min. of oxygen, but because of unavoidable leakage, it should be administered at a rate of 8 liters per min., if the oxygen concentration is to be kept at 60 per cent. CO_2 eliminated by the patient tends to accumulate within the tent, but a CO_2 concentration of 1.5 to 2 per cent is not harmful. It is important to keep the air within the tent dry and cool; otherwise the atmosphere easily becomes intolerably damp and hot. The air in the tent is made to circulate through a cooler containing ice or an electrically operated air-conditioning apparatus, where the temperature is lowered and the water breathed out by the patient is condensed.

FIG. 156. McKesson's oxygen tent. The receptacle beside the bed contains ice to cool and dry the air. Note the tubes for circulation of air between tent and cooler.

Oxygen can be administered efficiently by means of a catheter introduced into the nose, with its tip in the nasopharynx. Oxygen is first made to bubble in water and is then passed through the catheter at a rate of 4 to 6 liters per min. The inspired air contains approximately 60 per cent oxygen, and some patients find this procedure more comfortable than an oxygen tent.

Helium has been used as a therapeutic agent by Barach.[1] This gas is very light and noninflammable. Its atomic and molecular weight is 4, and its density compared with that of hydrogen (H_2) is 2. The density of nitrogen (molecular weight 28) is 14. Artificial air is made up of 21 per cent O_2 and 79 per cent helium. Helium, owing to its lightness, passes through the small bronchi and alveolar ducts much more easily than ordinary air. It has been used with good results in the treatment of asthma.

[1] BARACH, A. L., *J. A. M. A.*, **107**, 1273, 1936.

Carbon dioxide was first used in the treatment of anoxia by Yandell Henderson. A mixture of 95 per cent O_2 and 5 per cent CO_2, sometimes called "carbogen," is frequently used.

In normal subjects concentrations of CO_2 below 2 per cent in air with 20 per cent O_2 do not modify the respiratory rhythm. A small increase in CO_2 concentration increases the depth of breathing, which increases by 30 per cent when 2 per cent CO_2 is breathed. Concentrations of 4.5 to 5 per cent CO_2 cause some difficulty in breathing, sometimes nausea and malaise, and a slight increase in respiratory frequency. A concentration of 8.5 per cent produces dyspnea and arterial hypertension, which become very distressing after 15 to 20 min. but which rapidly disappear on breathing air. The upper limit tolerated lies between 7 and 9 per cent; higher concentrations (10 to 11 per cent) cause incoordination of movements and loss of consciousness. The response is even more marked with concentrations of 15 and 20 per cent, which are nevertheless tolerated up to 1 hr. by experimental animals. Concentrations of 25 to 30 per cent cause circulatory and respiratory depression, progressive loss of reflexes, coma, and death.

The reasons for administering CO_2 in ade-

quate concentrations are the following: Acute anoxia increases the excitability of the respiratory center reflexly by stimulating the chemoreceptors in the carotid and aortic bodies, but sufficiently intense or prolonged anoxia depresses the respiratory center and breathing becomes shallow. This is due to an exaggeration of the Hering-Breuer reflex, and anoxia is thus increased because of insufficient alveolar ventilation. At the same time there are serious disturbances in the tissues, especially in the myocardium and the central nervous system. Base migrates from the blood to the tissues. The importance of this acidosis of asphyxia has not been clearly established. If anoxia has not reached an advanced stage, artificial respiration with air or oxygen is sufficient to restore the patient to a normal condition, but if anoxia has been prolonged to a more serious stage, artificial respiration is no longer efficacious and the patient dies. If 95 per cent O_2 and 5 to 7 per cent CO_2 is used instead of air or oxygen, he is rapidly restored. A completely satisfactory explanation of the mechanism by which CO_2 acts in this case has not been given. Apparently the nerve centers (respiratory and vasomotor), severely depressed by anoxia, are no longer capable of responding to normal CO_2 concentrations; therefore to obtain a response the CO_2 concentration in alveolar air (and in the blood) should be raised. Pulmonary ventilation is then increased, the blood pressure rises, and the heartbeat is stronger.

The most evident effect on respiration of this treatment is the increase in its depth, which assures a more efficient alveolar ventilation and prevents or suppresses atelectasis by the wider expansion of the lung. This last effect is of importance in certain conditions in which atelectasis is a factor in pulmonary infection, as in postanesthetic atelectasis and the normal atelectasis of the newborn child. In normal babies this factor is not very important, but in weak or premature babies respiratory movements are inadequate and frequently do not expand the lungs completely. CO_2 inhalation is also useful in the shallow breathing of pneumonia, but it should not be administered when there is CO_2 retention as in stagnant anoxia.

Resuscitation. In severe anoxia and asphyxia the respiratory movements cease and serious disturbances in the circulation occur (tachycardia, hypotension, circulatory collapse). The response of the respiratory center is abnormal; it is less sensitive to chemical stimuli, and it can be depressed by substances which normally stimulate it. These respiratory and circulatory disturbances cause damage in the tissues most sensitive to anoxia (myocardium and central nervous system).

Resuscitation consists in rapidly taken measures to prevent irreversible damage to the tissues in cases of asphyxia or anoxia. These measures are artificial respiration, oxygen administration (preferably with 5 to 7 per cent CO_2), and treatment of circulatory collapse. If anoxia is not too severe, artificial respiration and oxygen (together with CO_2) will be enough to restore the patient. More advanced cases will need transfusion of blood or plasma, and the injection of vasoconstrictor drugs. The patient should be kept warm by adequate covering and heating.

ACCIDENTS DUE TO DECOMPRESSION

Accidents due to decompression were first observed in divers, in men working in caissons under water, and in sailors rescued from submarines. These accidents produce a syndrome known as "caisson disease." More recently the same effects have been observed in fighter pilots climbing rapidly to great heights.

A descent of 10 m. (32.8 ft.) below the surface in water causes an increase in pressure equal to 1 atmosphere. The air in the diver's helmet or in the caisson is renewed by pumping it in at a pressure varying with the depth and letting it out by a system of valves. The greatest depths so far reached are the equivalent of 10 atmospheres. The diver feels some discomfort on descending, such as pain in the ear owing to the difference in pressure on both sides of the eardrum, which is relieved by swallowing, thus equalizing the pressures; but serious disturbances occur only when he is returning to the surface, *i.e.*, on decompression.

While the diver is working under a high pressure, gases in the air are dissolved in the blood and tissues proportionately to the partial pressure of the gas, according to Henry's law. As he comes up to the surface, the pressure falls and the gases dissolved in the blood under pressure are released and form bubbles, as in an unstoppered soda-water bottle. Oxygen thus set free is rapidly absorbed by the cells, but nitrogen remains and can obstruct the finer vessels and capillaries (air embolism). These embolisms can

produce serious and even fatal damage when the heart or brain vessels are obstructed. The main symptoms are pain in the joints and bones, commonly known as the "bends"; paralysis caused by gaseous embolism in the spinal cord; and asphyxia and a choking sensation due to the accumulation of nitrogen bubbles in the large veins, the right heart cavities, and the pulmonary vessels.

To avoid accidents, decompression should be carried out gradually, *e.g.*, divers should rise to the surface very slowly. The same is true of a crew leaving a submerged submarine; the sailors are provided with masks and oxygen bags that permit them to remain submerged for some time. Nitrogen can then be gradually eliminated as the blood passes through the lung. When the men have been submitted to great pressures, sometimes several hours is needed to perform decompression safely. If signs or symptoms of caisson disease appear after the men have arrived at the surface, they should be immediately subjected to pressure in special tanks (Fig. 157), and then gradually decompressed. Aviators who must ascend rapidly to great heights are previously made to breathe pure oxygen so as to eliminate the nitrogen dissolved in the blood and tissues. The elimination of nitrogen in the gut in postoperative intestinal dilatation, or in the peritoneum or cerebral ventricles after air has been injected with the purpose of taking x-ray pictures of these cavities, can be hastened by breathing 95 per cent oxygen.[1] The partial pressure of nitrogen in the lungs is thus reduced to a negligible figure, and N is rapidly eliminated from the blood; approximately 60 per cent is lost in the first hour.

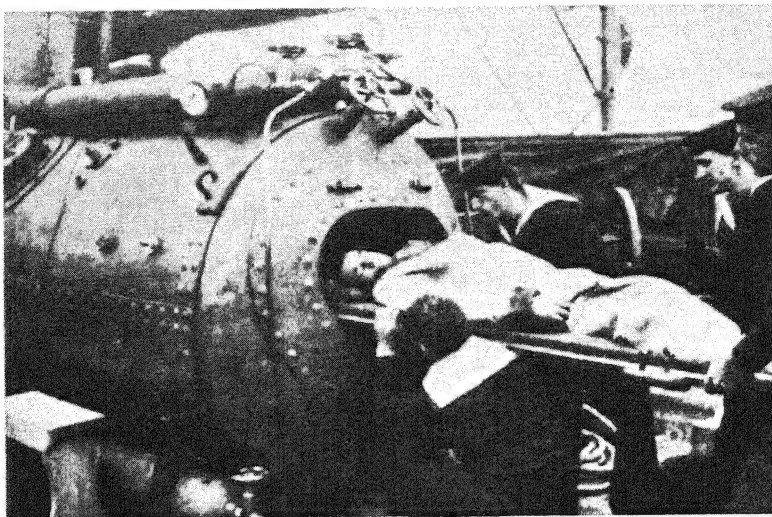



FIG. 157. Decompression tank. A patient is being introduced into the tank, in which the pressure will be raised and then very slowly lowered. (*Haldane, J. S., and J. G. Priestly, "Respiration," Oxford University Press, New York, 1935, p. 336.*)

BIBLIOGRAPHY

BARCROFT, J., "The Respiratory Function of the Blood," Part 1, Lessons from High Altitudes, Cambridge, New York, 1925.

BEAN, J. W., Effects of Oxygen at Increased Pressure, *Physiol. Rev.*, **25**, 1–147, 1945.

BEHNKE, A. R., Decompression Sickness Incident to Deep Sea Diving and High Altitude Ascent, *Medicine*, **24**, 381, 1945.

FULTON, J. F., *et al.*, "Decompression Sickness, Divers' and Fliers' Bends and Related Syndromes," Saunders, Philadelphia, 1951.

HENDERSON, Y., "Adventures in Respiration," Williams & Wilkins, Baltimore, 1938.

HURTADO, A., Chronic Mountain Sickness, *J. A. M. A.*, **120**, 1278, 1942.

LUNDSGAARD, C., and D. D. VAN SLYKE, Cyanosis, *Medicine*, **2**, 1, 1923.

NESI, J. A., Encefalopatía Anoxémica Postanestésica, *Rev. Asoc. méd. argent.*, **58**, 816, 1944.

RICHARDS, D. W., and A. L. BARACH, Prolonged Residence in High Oxygen Atmospheres. Effects on Abnormal Individuals and on Patients with Chronic Cardiac and Pulmonary Insufficiency, *Quart. J. Med.*, **3**, 437, 1934.

VAN LIERE, E. J., "Anoxia, Its Effects on the Body," University of Chicago Press, Chicago, 1942.

[1] FINE, J., B. M. BANKS, and L. HERMANSON, *Ann. Surg.*, **103**, 375, 1936; SAKLAD, M., and A. M. BURGESS, *J. A. M. A.*, **124**, 831, 1944.

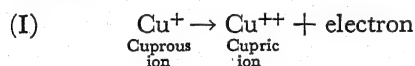
CHAPTER 34

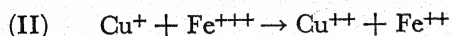
Cell Respiration

OXYGEN COMBINES IN the tissues with several substances with the formation of CO_2 and water. Animals obtain from these oxidation processes the greater part of the energy needed for their bodily activities.

The significance of the term "oxidation" has gradually changed. Originally it was used for processes in which there is fixation of oxygen; later, for processes in which electrons are lost. Usually the following processes are considered as oxidations: (*a*) loss of electrons; (*b*) loss of hydrogen atoms; (*c*) addition of oxygen. Inverse processes are considered as reductions. At first sight it would seem that widely different reactions are grouped under the term oxidation; nevertheless these processes have much in common, and it will be seen how reaction (*c*) can provoke processes (*a*) and (*b*).

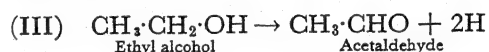
Loss of electrons. A typical example of this process is the conversion of cuprous to cupric ion, which can be expressed in a simplified way as follows:

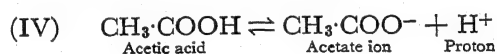
$$\text{(I)} \qquad \underset{\text{Cuprous ion}}{Cu^{+}} \rightarrow \underset{\text{Cupric ion}}{Cu^{++}} + \text{electron}$$

A solution of cuprous ions can be obtained by dissolving cuprous chloride in water. The salt dissociates into Cl^- and Cu^+, but as Cl^- is not modified in the process here considered, it has been omitted in Equation I. Cuprous ions lose an electron when a substance is added to the solution which has a greater affinity for electrons than these ions, *i.e.*, an electron acceptor. For example, ferric chloride has this effect. When dissolved this salt dissociates into Cl^- and ferric ions, and the following reaction takes place:

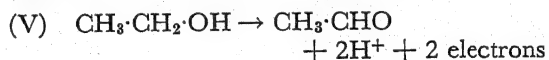
$$\text{(II)} \qquad Cu^{+} + Fe^{+++} \rightarrow Cu^{++} + Fe^{++}$$

One electron passes from Cu^+ to Fe^{+++}, cuprous ions are oxidized to cupric ions, and ferric ions are reduced to ferrous ions.

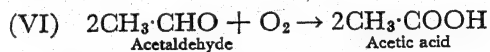
A change of valence takes place in this process; there is an increase of one positive valence in oxidation and a decrease in reduction. There is also a change of valence when iodide (I^-) or chloride (Cl^-) ions lose electrons and are oxidized to iodine (I_2) or chlorine (Cl_2). In this case there is a decrease of one negative valence.

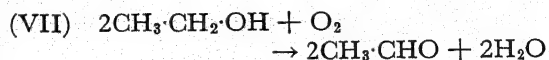
Loss of hydrogen atoms or dehydrogenation. This type of oxidation is illustrated by the following equation:

$$\text{(III)} \qquad \underset{\text{Ethyl alcohol}}{CH_3{\cdot}CH_2{\cdot}OH} \rightarrow \underset{\text{Acetaldehyde}}{CH_3{\cdot}CHO} + 2H$$

In this case hydrogen atoms are lost, a process that is not the same as the loss of protons that occurs on dissociation of an acid:

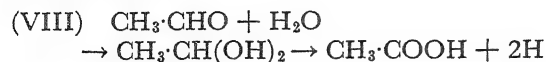
$$\text{(IV)} \qquad \underset{\text{Acetic acid}}{CH_3{\cdot}COOH} \rightleftharpoons \underset{\text{Acetate ion}}{CH_3{\cdot}COO^-} + \underset{\text{Proton}}{H^+}$$

The only difference between the processes of dehydrogenation and loss of electrons is that in the former the electron lost is accompanied by a proton. Equation III could be expressed as a loss of electrons in the following way:

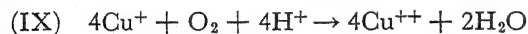
$$\text{(V)} \qquad CH_3{\cdot}CH_2{\cdot}OH \rightarrow CH_3{\cdot}CHO + 2H^+ + 2 \text{ electrons}$$

Addition of oxygen. Oxidation of acetaldehyde into acetic acid can be taken as an example:

$$\text{(VI)} \qquad \underset{\text{Acetaldehyde}}{2CH_3{\cdot}CHO} + O_2 \rightarrow \underset{\text{Acetic acid}}{2CH_3{\cdot}COOH}$$

Dehydrogenation of ethyl alcohol (Equation III) could also be expressed as oxidation by oxygen:

$$\text{(VII)} \qquad 2CH_3{\cdot}CH_2{\cdot}OH + O_2 \rightarrow 2CH_3{\cdot}CHO + 2H_2O$$

Oxygen acts here as a hydrogen acceptor (oxidizing agent) and at the same time is reduced (accepts hydrogen) and forms water. Ethyl alcohol is dehydrogenated (oxidized) and gives acetaldehyde.

Oxidation of acetaldehyde by oxygen (Equation VI) can also be expressed as a dehydrogenation by postulating the formation first of aldehyde hydrate:

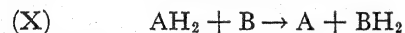
$$\text{(VIII)} \quad CH_3{\cdot}CHO + H_2O \rightarrow CH_3{\cdot}CH(OH)_2 \rightarrow CH_3{\cdot}COOH + 2H$$

Finally, the essential unity of all oxidation processes is made evident on considering autooxidation of cuprous ions. Aqueous solutions of cuprous ions absorb oxygen spontaneously according to the following equation:

$$\text{(IX)} \quad 4Cu^{+} + O_2 + 4H^{+} \rightarrow 4Cu^{++} + 2H_2O$$

The protons taking part in the reaction arise from the dissociation of water: $H_2O \rightleftharpoons H^{+} + OH^{-}$. The oxidizing agent is oxygen, which is reduced when water is formed. It could also be said that oxygen acts as a hydrogen acceptor, or that in the presence of protons it accepts electrons.

Since loss of electrons or hydrogen atoms and addition of oxygen are only different aspects of oxidation processes, these terms will be used indiscriminately in the following discussion.

THE MECHANISM OF CELL RESPIRATION

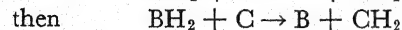
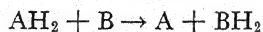
Oxygen is the final oxidizing agent in animal tissues, but there are many intermediary reactions in which other substances act as oxidizing agents, *i.e.*, as hydrogen or electron acceptors. Oxidation-reduction between two substances A and B can be generally expressed as follows:

$$\text{(X)} \quad AH_2 + B \rightarrow A + BH_2$$

B can be oxygen or any other oxidizing substance. A is oxidized and B is reduced.

Substances on which enzymes act are known as substrates. Between a substrate that is oxidized and oxygen, there are intermediary substances that are oxidized and reduced reversibly in the course of the process. If the substrate is represented by AH_2 the first reaction is

$$AH_2 + B \rightarrow A + BH_2$$

then

$$BH_2 + C \rightarrow B + CH_2$$

and so on, until finally

$$2ZH_2 + O_2 \rightarrow 2Z + 2H_2O$$

Substances B, C, D, . . . , Z act together with specific proteins. In some cases these substances are firmly bound to the respective proteins, and they are not easily dissociated. The substance is then called the prosthetic group of the enzyme. If, on the contrary, the complex is easily dissociated, the substance is called a coenzyme. There is no fundamental difference between prosthetic groups and coenzymes, only differences in degree; therefore it might be better to consider these terms synonymous.

Enzymes are proteins, *i.e.*, they are formed by the union of amino acids. The prosthetic groups or coenzymes are of a different structure, and they can be used to classify oxidation-reduction enzymes. The principal enzymes that take part in cell respiration are (*a*) those which have pyridine-nucleotides as coenzymes; (*b*) flavoproteins; (*c*) cytochromes, which have a prosthetic group similar to heme in hemoglobin; (*d*) other enzymes that will be considered later.

Dehydrogenases. Pyridine-nucleotides. Enzymes that act directly in the oxidation of the substrate are usually called "dehydrogenases." Typical examples of these enzymes are those in which activity is dependent on the presence of pyridine-nucleotides.

The term "pyridine-nucleotide" is due to the fact these substances contain nicotinic amide, which has a pyridine ring. Nicotinic amide united to ribose and phosphoric acid forms a nucleotide.

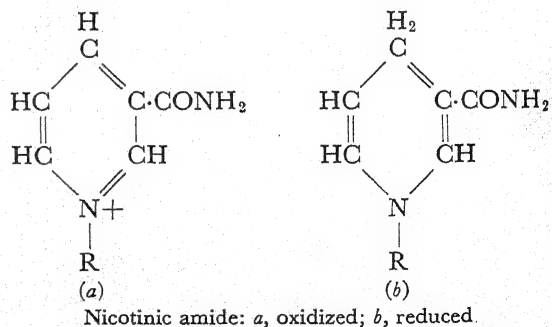
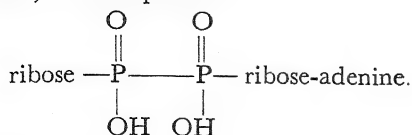
Two pyridine-nucleotides are known. One of them, diphosphopyridine-nucleotide (DPN), also called "coenzyme I" or "cozymase," is formed by the union of the following molecules: 1 nicotinic amide, 1 adenine, 2 ribose, and 2 phosphoric acid; it is therefore a dinucleotide.

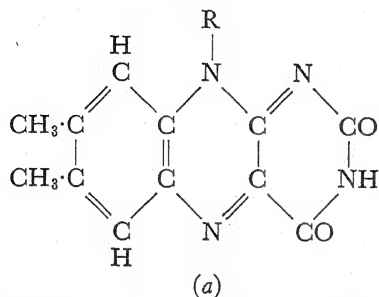
The other known pyridine-nucleotide is triphosphopyridine-nucleotide (TPN) which has three phosphate groups instead of two.

```
        H                          H2
        C                          C
      /   \\                     /   \\
   HC       C·CONH2          HC       C·CONH2
   ||       |                ||       ||
   HC       CH               HC       CH
      \   //                    \   /
        N+                         N
        |                          |
        R                          R
       (a)                        (b)
```

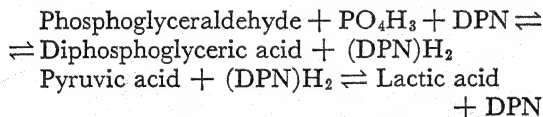
Nicotinic amide: *a*, oxidized; *b*, reduced

In DPN, R corresponds to

```
        O      O
        ||     ||
ribose —P——————P— ribose-adenine.
        |      |
        OH     OH
```

In TPN there is an additional phosphate group.

Dehydrogenases catalyze the reaction between the substrate and DPN or TPN. The substrate is oxidized, and the nicotinic amide part of the nucleotide is reduced. This is a reversible reaction, and for a substrate AH_2, it can be diagrammatically expressed as follows:

$$AH_2 + (DPN) \rightleftharpoons (DPN)H_2 + A$$

These coenzymes have been found in all living organisms, and they take part in reactions of fundamental importance. On considering carbohydrate metabolism (Chap. 41) it will be seen how in the disintegration of glycogen, or glucose, to lactic acid, there are a series of reactions in two of which DPN takes part. These reactions are the following:

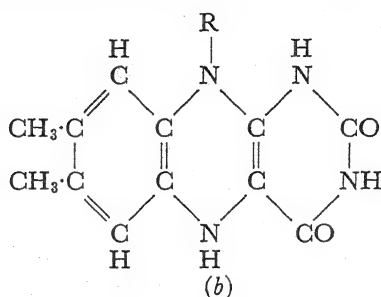
$$\text{Phosphoglyceraldehyde} + PO_4H_3 + DPN \rightleftharpoons$$
$$\rightleftharpoons \text{Diphosphoglyceric acid} + (DPN)H_2$$
$$\text{Pyruvic acid} + (DPN)H_2 \rightleftharpoons \text{Lactic acid} + DPN$$

In the anaerobic process DPN acts as an intermediary between the two reactions; it is reduced in the former and oxidized in the latter. In the presence of oxygen this does not take place because there are mechanisms that permit the oxidation of reduced DPN. In this case, oxygen instead of pyruvic acid is the final oxidizing agent.

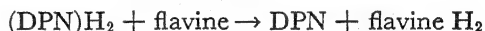
Pyridine-nucleotides are not oxidized directly by oxygen, but by other enzymes and their prosthetic groups.

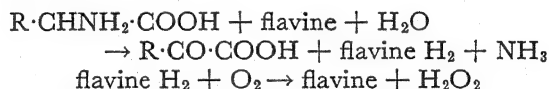
In addition to the phosphoglyceraldehyde and lactic dehydrogenases mentioned above, there are others that also act with DPN. These are malic, β-hydroxybutyric, and alcohol dehydrogenases. Others act in conjunction with TPN, such as glucose-6-phosphate, phosphohexonic, isocitric dehydrogenases, etc.

Flavoproteins. Pyridine-nucleotides can be reoxidized by reacting with flavoproteins, the prosthetic group of which, *i.e.*, flavine-dinucleotide, is a derivative of riboflavin (vitamin B_2).

```
            R                                  R
      H     |                           H      |     H
      C     N     N                     C      N     N
    //  \  / \  // \                  //  \  /  \  /  \
CH3·C    C     C     CO           CH3·C    C      C     CO
    |    ||    |     |                |    ||     ||    |
CH3·C    C     C     NH           CH3·C    C      C     NH
    \\  / \  // \   /                 \\  /  \   /  \  /
      C     N     CO                    C      N      CO
      H                                 H      H
          (a)                                 (b)
```

Flavine-nucleotide: *a*, oxidized; *b*, reduced. R is ribityl-(phosphate)$_2$ ribose-adenine in the flavine-dinucleotide, and ribityl-phosphate in the flavine-mononucleotide.

As in other processes, the pyridine-nucleotide is oxidized and the flavine is reduced.

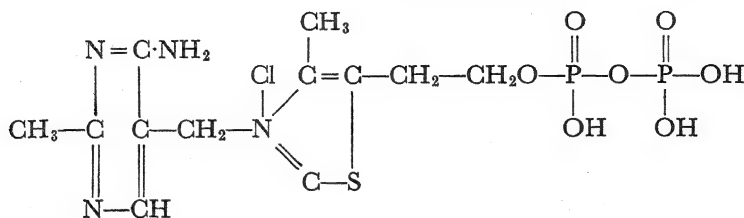
$$(DPN)H_2 + \text{flavine} \rightarrow DPN + \text{flavine } H_2$$

Another enzyme of the flavoprotein type oxidizes D-amino acids:

$$R{\cdot}CHNH_2{\cdot}COOH + \text{flavine} + H_2O$$
$$\rightarrow R{\cdot}CO{\cdot}COOH + \text{flavine } H_2 + NH_3$$
$$\text{flavine } H_2 + O_2 \rightarrow \text{flavine} + H_2O_2$$

In this case the flavoprotein is reoxidized directly by oxygen. Enzymes of this type, acting as the only link between the substrate and oxygen, are called oxidases. The above equation represents the activity of D-amino acid oxidase.

Other flavoproteins are not reoxidized directly by oxygen. *In vitro* they can be oxidized by several reagents, *e.g.*, methylene blue. *In vivo* they are reoxidized by other enzymes, the cytochromes.

Cytochromes. The prosthetic group of cytochromes is an iron-porphyrin similar to heme in hemoglobin. Several cytochromes (a_1, a_3, b, c) have been described. The best known is cytochrome c, which has been obtained pure. Other cytochromes have been identified by their absorption spectra, *i.e.*, the property of absorbing light of a given wavelength. Each one of the cytochromes has a typical absorption spectrum that varies according to whether it is in the oxidized or reduced state.

The catalytic effect of the cytochromes on cell oxidations is due to the capacity of their prosthetic group to be oxidized or reduced. They can thus form links in the chain of catalyzers that act in the oxidation of substrates by molecular oxygen.

```
                         CH3                          O      O
       N=C·NH2            |                           ||     ||
       |    |       Cl   C=C—CH2—CH2O—P—O—P—OH
       |    |          | /    |                       |      |
CH3—C    C—CH2—N        |                      OH     OH
       ||   ||          \     |
       ||   ||           C—S
       N—CH
```

Diphosphothiamine.

Oxidized cytochromes have ferric ion (Fe^{+++}) in the porphyrine, which is reduced to ferrous ion (Fe^{++}) in reduced cytochrome. The difference is similar to that between hemoglobin, which has its iron as Fe^{++}, and methemoglobin, which has its iron as Fe^{+++}. It should be noted that hemoglobin combined with oxygen is not generally considered as oxidized hemoglobin. The combination is, in this case, easily reversible; therefore the term "oxygenated hemoglobin" or "oxyhemoglobin" is used to differentiate HbO_2 from truly oxidized hemoglobin, which is methemoglobin.

The mechanism by which cytochromes react with oxygen is not known in detail; cytochrome c is not autooxidizable, *i.e.*, a solution of reduced cytochrome is not oxidized by molecular oxygen.

Cytochrome oxidation is carried out by cytochrome oxidase, Warburg's *Atmungsferment* (German *Atmung*, respiration) or respiratory enzyme. This enzyme is not yet well known; perhaps it is an autooxidizable cytochrome that can oxidize other cytochromes.

A typical reaction of cytochrome oxidase is its combination with cyanide. A very small concentration of cyanide inhibits cytochrome oxidase and almost completely suppresses oxygen consumption in the tissues.

The present knowledge of tissue oxidations is far from being complete, but apparently some substrates are oxidized by a chain of enzymes and coenzymes or prosthetic groups, more or less as follows:

Substrate → dehydrogenase + DPN or
TPN → flavoprotein → cytochrome → O_2

Coenzymes and prosthetic groups. It is useful to recapitulate here some of the facts already mentioned, adding others which have not yet been considered.

Many enzymes are made up of two parts: (*a*) a protein, in the strict sense of the term, *i.e.*, formed only of amino acids; (*b*) a nonprotein moiety. If the nonprotein portion is easily separated from the protein it is called the coenzyme. If, on the contrary, both parts are firmly united the nonprotein part is known as the prosthetic group.

Pyridine-nucleotides take part in the activity of some dehydrogenases; they are nicotinic amide derivatives. Flavoproteins have a derivative of riboflavin as prosthetic group. To this group belong D-amino acid oxidase, xanthinoxidase, and the flavoproteins that mediate in the oxido-reduction between pyridine-nucleotides and cytochromes.

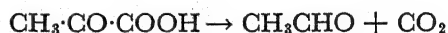
The prosthetic group of cytochromes is an iron-porphyrin similar to heme in hemoglobin.

Catalase also has an iron-porphyrin as a prosthetic group. This enzyme catalyzes the decomposition of hydrogen peroxide and liberates molecular oxygen:

$$2H_2O_2 \rightarrow O_2 + 2H_2O$$

Thiamine combined to two molecules of phosphoric acid forms diphosphothiamine, the coenzyme involved in the transformation of pyruvic acid.

In brewer's yeast diphosphothiamine is the coenzyme of carboxylase, which catalyzes decarboxylation of pyruvic acid as follows:

$$CH_3{\cdot}CO{\cdot}COOH \rightarrow CH_3CHO + CO_2$$

Owing to this activity diphosphothiamine is also known as cocarboxylase.

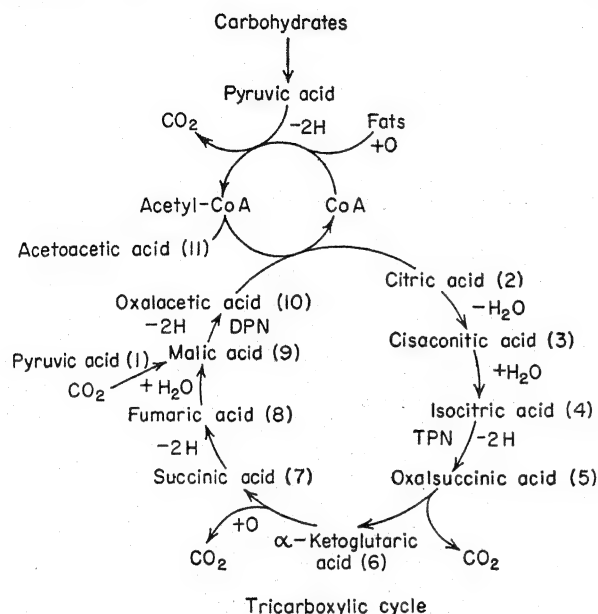
In animal tissues the reaction is different; decarboxylation and oxidation of pyruvic acid take place, and the products of these processes are further oxidized by way of the tricarboxylic cycle. In B_1 vitamin deficiency, there is an insufficient amount of diphosphothiamine; oxidation

of pyruvic acid is therefore delayed, and it is accumulated in abnormal amounts.

A pyridoxine derivative (pyridoxal phosphate) is the coenzyme of certain amino-acid decarboxylases in bacteria and in animal tissues of DOPA decarboxylase (DOPA = dioxyphenylalanine).

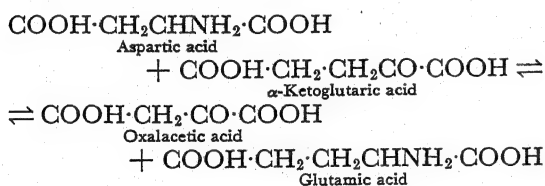
thioethanolamine forms coenzyme A, which takes part in certain acetylation processes. This enzyme is active in the acetylation of choline with formation of acetylcholine, which is of great physiologic importance. Moreover it also seems to play a part in the formation of citric acid (see diagram of tricarboxylic cycle).

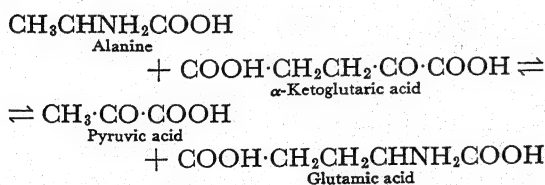
Carbohydrates
Pyruvic acid
CO_2 −2H Fats +O
Acetyl-Co A
Co A
Acetoacetic acid (11)
Citric acid (2)
$-H_2O$
Oxalacetic acid (10)
−2H DPN
Cisaconitic acid (3)
Pyruvic acid (1) Malic acid (9)
$+H_2O$
CO_2 $+H_2O$
Fumaric acid (8)
Isocitric acid (4)
−2H
TPN −2H
Succinic acid (7)
Oxalsuccinic acid (5)
+O
α−Ketoglutaric acid (6)
CO_2 CO_2

Tricarboxylic cycle

All reactions are reversible, but the arrows have been depicted only in one direction, in order to simplify the diagram.

1) $CH_2-CO.COOH$
2) $COOH-CH_2-COH(COOH)-CH_2-COOH$
3) $COOH-CH_2-C(COOH)=CH-COOH$
4) $COOH-CH_2-CH(COOH)-CHOH-COOH$
5) $COOH-CH_2-CH(COOH)-CO-COOH$
6) $COOH-CH_2-CH_2-CO-COOH$
7) $COOH-CH_2-CH_2-COOH$
8) $COOH-CH=CH-COOH$
9) $COOH-CHOH-CH_2-COOH$
10) $COOH-CO-CH_2-COOH$
11) $CH_3-CO-CH_2-COOH$

Pyridoxal phosphate is also the coenzyme of transaminases. Aspartic-transaminase catalyzes the following reaction:

$$\underset{\text{Aspartic acid}}{COOH{\cdot}CH_2CHNH_2{\cdot}COOH} + \underset{\alpha\text{-Ketoglutaric acid}}{COOH{\cdot}CH_2{\cdot}CH_2CO{\cdot}COOH} \rightleftharpoons \underset{\text{Oxalacetic acid}}{COOH{\cdot}CH_2{\cdot}CO{\cdot}COOH} + \underset{\text{Glutamic acid}}{COOH{\cdot}CH_2{\cdot}CH_2CHNH_2{\cdot}COOH}$$

Alanine-transaminase catalyzes the following reaction:

$$\underset{\text{Alanine}}{CH_3CHNH_2COOH} + \underset{\alpha\text{-Ketoglutaric acid}}{COOH{\cdot}CH_2CH_2{\cdot}CO{\cdot}COOH} \rightleftharpoons \underset{\text{Pyruvic acid}}{CH_3{\cdot}CO{\cdot}COOH} + \underset{\text{Glutamic acid}}{COOH{\cdot}CH_2CH_2CHNH_2COOH}$$

There are similar enzymes for many other amino acids. Pantothenic acid (see Chap. 49, Vitamins) combined with adenosine, phosphates, and

These facts are summarized in Table 30.

The prosthetic groups of several enzymes are still unknown, *e.g.*, that of succinodehydrogenase, which catalyzes the oxidation of succinic acid to fumaric acid, one of the reactions in the tricarboxylic cycle. The prosthetic groups of monoamino oxidase and histaminase are also unknown.

Although many coenzymes derive from vitamins, others do not, *e.g.*, glucose-1,6-diphosphate which acts as the coenzyme in the conversion of glucose-1-phosphate to glucose-6-phosphate.

The tricarboxylic cycle. Pyruvic acid formed in the disintegration of carbohydrate and the two-carbon-atom units resulting from β-oxidation of fatty acids are metabolized in a complex cyclic process, known as the tricarboxylic cycle. The name comes from the fact that some of the intermediate products, such as isocitric and aconitic acids, have three carboxyl groups.

Oxidation of pyruvic acid results in the formation of acetyl-coenzyme A. In this process a

thermostable factor, diphosphothiamine, takes part.

The cycle begins by the condensation of acetyl-coenzyme A and oxalacetic acid with the formation of citric acid (see diagram). A series of reactions leads to the formation of isocitric acid, then α-ketoglutaric acid, succinic acid, etc., and finally oxalacetic acid is again formed. In each turn of the cycle, two molecules of CO_2 are set free.

Table 30. Vitamins in Enzyme Systems

Vitamin	*Coenzyme or prosthetic group*	*Enzyme*
Nicotinic amide (PP).	Diphosphopyridine-nucleotide and triphosphopyridine-nucleotide	Several dehydrogenases
Riboflavin (B_2)......	Flavine-nucleotides	Flavoproteins, *d*-amino-acid oxidases, xanthine oxidase, etc.
Thiamine (B_1).......	Diphosphothiamine	Enzymes acting on pyruvic acid
Pyridoxine (B_6)......	Pyridoxal phosphate	Transaminases, amino-acid decarboxylases, etc.
Pantothenic acid.....	Coenzyme A	Acetylation enzymes
———	Iron-porphyrins	Cytochromes

The following reactions in the cycle are oxidations: (*a*) the conversion of isocitric acid into oxalsuccinic acid, in which TPN takes part; (*b*) the conversion of malic acid into oxalacetic acid, in which DPN takes part; (*c*) the oxidation of α-ketoglutaric acid and of succinic acid, in which the coenzymes or prosthetic groups are still unknown.

Some of the details in the tricarboxylic cycle still need to be confirmed, but the general outline of the cycle has been solidly established on experimental data. This cycle also explains the interrelation between carbohydrate and fat metabolism, since both have a common path of oxidation. Moreover, some of the amino acids can give rise to certain substances in the cycle. For example, aspartic and glutamic acids can be converted by transamination into oxalacetic and α-ketoglutaric acids. The tricarboxylic cycle, therefore, is a common path in the metabolism of carbohydrates, fats, and proteins.

METHODS OF STUDY

Chemical reactions that take place in the whole animal can be studied by the analysis of substances entering the organism (food, inspired air) and the end products of metabolism in the expired air, urine, and feces. Chemical reactions can be observed in isolated organs perfused with an adequate nutrient fluid, or in slices of surviving tissue, submerged in blood serum or in Ringer's solution. These slices must be extremely thin, so that oxygen and the substances in the nutrient fluid can penetrate rapidly into the tissue; if the slices are too thick the deeper parts are deprived of oxygen and food. The consumption of substances added to the fluid can thus be quantitatively studied by chemical methods; also the oxygen consumption and CO_2 production can be accurately determined. Several microrespirometers have been designed for this purpose. Warburg's is the one most frequently used.

Warburg's apparatus. (Fig. 158.) This consists of a capillary manometer for measuring the changes in pressure occurring in the flask with which it is connected. This flask is submerged in a bath maintained at constant temperature; within it the substance to be studied is placed (tissue slice, minced tissue, enzyme in solution). If only the oxygen consumption is to be determined, a few drops of NaOH are placed in a small central compartment; CO_2 formed is immediately absorbed by the NaOH, so that the changes in pressure are due solely to changes in the amount of oxygen within the flask. It is then easy to calculate, from the fall in pressure registered by the manometer, the amount (in cubic millimeters) of oxygen consumed.

Thunberg's method. Thunberg's tube is a test tube stoppered in such a way that gases can be evacuated and an airtight closure can be made subsequently. In this tube the speed of reduction of methylene blue or similar dyes can be measured. If a solution of an enzyme together with a substrate and methylene blue is placed in a Thunberg tube and the substrate is oxidized, methylene blue will be reduced. On reduction, the dye is converted into its colorless leukobase; therefore the velocity of decoloration

of the dye is a measure of the velocity of oxidation of the substrate. Moreover this velocity is proportional to the enzyme concentration. A vacuum must first be made within the tube, because the leukobase is autooxidizable, so that if oxygen were present the dye would be reoxidized continuously and would never lose its color.

Inhibitors. Many of the intermediary reactions take place one after the other at approximately equal velocities; therefore the intermediary substrates do not accumulate in appreciable amounts and cannot be measured or isolated. To overcome this difficulty inhibitors that act specifically on one of the reactions only are added to the system; *e.g.*, if fluoride is added, enolase is inhibited and phosphoglyceric acid accumulates. Other examples of more or less specific inhibitors are cyanide, which acts on cytochrome oxidase; monoiodacetic acid, which inhibits phosphoglyceraldehyde dehydrogenase; etc.

Isolation of enzymes. The extraction of minced tissues with water or saline solution yields mixtures of enzymes contained in the tissues. Each one of the enzymes must then be separated from the others. This can be accomplished in some cases by fractional precipitation with salts, or by organic solvents, or by adsorption. In other cases a more resistant enzyme is separated from others less resistant by heat, acidification, or treatment with alcohol or acetone.

The study of intermediary carbohydrate metabolism is a good example of the value of this method. At first tissue extracts that converted glycogen into lactic acid were obtained. Then the enzymes were separated one by one, and the products of each reaction were isolated.

Spectrophotometric methods. By measuring the absorption of visible or ultraviolet light it is possible not only to identify some of the components in the oxidation-reduction enzyme system, but also to measure the velocity of the reactions.

Reduced pyridine-nucleotides show an absorption band at a wavelength of 345 mμ, (1 mμ = 10 A). The oxidized form does not show this absorption band, therefore it is possible to measure the speed of reduction by the rate of absorption at 345 μm.

Flavines have a yellow color; their reduction can be followed by the naked eye because they lose color. Each one of the cytochromes has typical absorption bands, which change on reduction. For example, with suspensions of brewer's yeast it is easy to observe the absorption band corresponding to reduced cytochrome c at 550 mμ. This band is visible in yeast under anaerobic conditions; but if the suspension is merely shaken in air the absorption band disappears owing to the reoxidation of cytochrome.

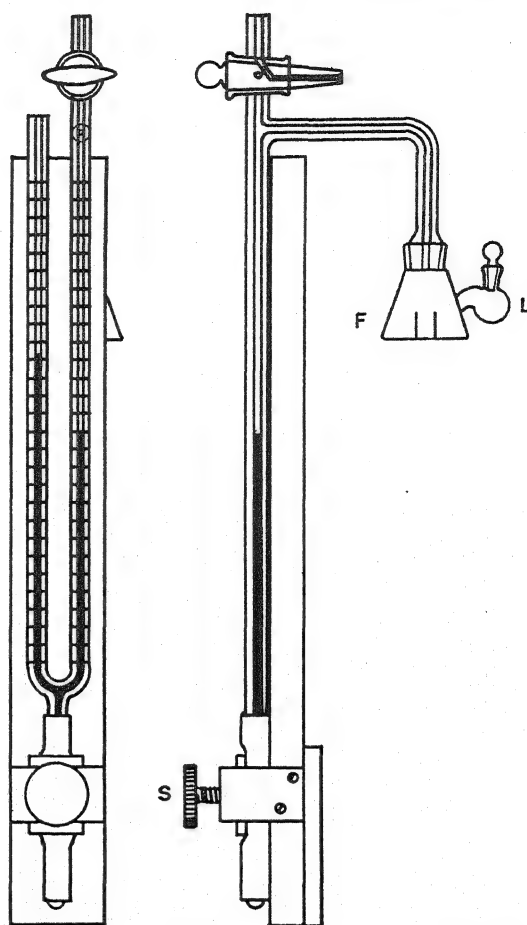



FIG. 158. Diagram of Warburg's apparatus. The manometer is a U-shaped tube, filled with a fluid of known density. By means of the screw *S* the fluid is kept at a certain level in the branch of the manometer that does not communicate with the flask, *F*; thus the level of the fluid in the other branch of the manometer measures the pressure in the flask. The reactions take place within *F*, which is submerged in a constant-temperature bath. Solutions can be placed in the lateral compartment *L* and added to the substrate without opening the system. The central tube in *F* is filled with alkali to absorb CO_2 when oxygen consumption is being measured.

BIBLIOGRAPHY

Advances in Enzymology (several papers).

Annual Review of Biochemistry (several articles on oxidation-reduction).

BALDWIN, E., "Dynamic Aspects of Biochemistry," 2d ed., Cambridge, New York, 1952.

DIXON, M., "Multiple-enzyme Systems," Cambridge, New York, 1949.

ELLIOT, K. A. C., Intermediary Metabolites and Respiratory Catalysis, *Physiol. Rev.*, 21, 267, 1941.

GREEN, D. E., "Mechanism of Biological Oxidations," Cambridge, New York, 1940.

GREEN, D. E., "Currents in Biochemical Research," Interscience Publishers, Inc., New York, 1946.

GUZMÁN BARRÓN, E. S., Cellular Oxidation Systems, *Physiol. Rev.*, **19**, 184, 1939.

KEILIN, D., Le Mécanisme de la respiration intracéllulaire, *Bull. Soc. chim. biol.*, **18**, 96, 1936.

LARDY, H. A., "Respiratory Enzymes," Burgess Publishing Co., Minneapolis, 1949.

OPPENHEIMER, C., and K. G. STERN, "Biological Oxidation," Nordemann Publishing Co., New York, 1939.

SUMNER, J. B., and K. MYRBÄCK, "The Enzymes. Chemistry and Mechanism of Action," Academic Press, Inc., New York, 1950–1952.

SUMNER, J. B., and G. F. SOMERS, "Chemistry and Methods of Enzymes," Academic Press, Inc., New York, 1947.

"Symposium on Respiratory Enzymes," University of Wisconsin Press, Madison, 1942.

TAUBER, H., "The Chemistry and Technology of Enzymes," Wiley, New York, 1949.

UMBREIT, W. W., R. H. BURRIS, and J. F. STAUFFER, "Manometric Techniques and Tissue Metabolism," 2d ed., Burgess Publishing Co., Minneapolis, 1949.

WURMSER, R., "Oxydations et réductions," Presses Universitaires de France, Paris, 1930.

SECTION FOUR

Digestion

CHAPTER 35

The General Physiology of Digestion

MAN, LIKE OTHER animals, cannot live without food. Substances that provide the specific mass and energy of living beings, and also those needed to maintain normal metabolism, are called foods. Foodstuffs can be classified (*a*) by their origin, as animal (meat, milk, eggs, fat), vegetable (leguminous plants, cereals, sugars, oil), and mineral (salts, water, oxygen); (*b*) by their chemical nature, as proteins, fats, carbohydrates, salts, water, etc.; (*c*) by their principal function, as energy suppliers (carbohydrates, fats, proteins, alcohol), tissue builders (protein, fats, carbohydrates, and some of the salts), and regulators of metabolism (vitamins, salts, water).

Foodstuffs must be altered before they can be used by the organism. In higher animals this alteration takes place in the digestive tract. In the process of digestion, complex, insoluble substances are changed into simpler, soluble, and diffusible forms that can be absorbed and assimilated. This process is performed by the action of the digestive juices secreted by the glands that pour their secretions into the digestive tract. These juices contain enzymes, which reduce the varied foodstuffs into a few simple substances that can be absorbed. Carbohydrates are transformed principally into glucose, fructose, and galactose; fats into fatty acids and glycerol; and proteins into simple amino acids. Water and salts pass through the digestive tract and are absorbed without any appreciable change.

Foodstuffs progress along the digestive tract and are mixed with the digestive juices by the movements of the different segments.

The digestive tract is formed by two or three muscular layers, lined internally by a mucosa, and covered, in parts, by a serous membrane. The mucosa has its origin in the endoderm. In the mouth the mucosa is a squamous stratified epithelium. The lower pharynx and the esophagus are lined by stratified pavemented epithelium. On the gastric side of the cardia there is an abrupt change to simple cylindrical epithelium, which lines the stomach and small and large intestines down to about 2 cm. from the anus, where the epithelium again becomes stratified and pavemented.

The cells of the mucosa lie on the lamina propria of the chorion. In most parts of the digestive tract there is a muscularis mucosae below the chorion, separated from the muscular layers by relatively lax connective tissue (submucosa). The mucosa of the small intestine is folded into villi, and the surface of absorption is thus increased considerably.

The digestive glands. The structure of all the digestive glands follows a similar pattern. The cells of the secretory epithelium form tubes or vesicles of different length and complexity, which communicate, by means of ducts of variable length, with the lumen of the segment of the tract from which they arise. Some of the digestive glands, *e.g.*, the liver, pancreas, and some of the salivary glands, have a single type of secretory cell, but most of them have several types. The glands of the intestinal crypts have at least four kinds of specialized cells; those of the stomach, three; and some of the salivary glands, two. The different types of cells are frequently disposed in separate layers; those that produce a fluid secretion occupy a peripheral situation, and the internal cells secrete a viscous juice rich in organic substances (proteins, enzymes). The more fluid secretion of the external cells can thus flush into the lumen of the gland the secretion of the central cells.

The digestive glands secrete about 5 liters of different juices in 24 hr.; most of this is reabsorbed by the intestine. The greater part of the substances forming this large mass comes from the blood.

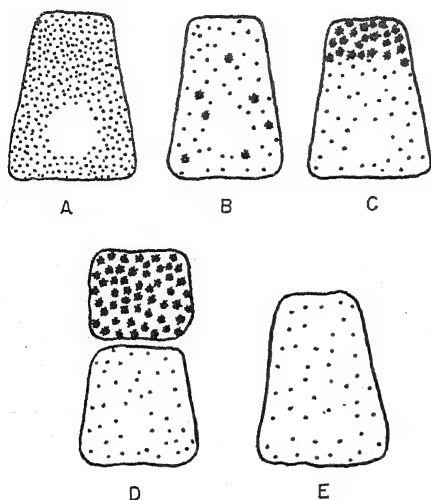



Fig. 159. Diagram of five stages in the process of secretion in a cell of the cat's submaxillary gland. *a*, resting cell; the dots represent the granules, the empty space represents the location of the nucleus. *b*, beginning of stimulation of the chorda tympani; the small granules join to form large ones. *c*, stimulation continues; the large granules collect at the distal end of the cell. *d*, stimulation continues; the granules are discharged together with the distal end of the cell. *e*, end of stimulation; the distal end of the cell is restored, and few granules remain. (*Holzlöhner, E., and C. Niessing, Ztschr. f. Biol., vol. 97, p. 563, 1936.*)

The cells of the liver are in direct contact with the blood vessels. The cells of other glands obtain water and crystalloids from the interstitial fluid surrounding the alveoli and tubes. The process of secretion is conditioned, therefore, by the permeability of the capillaries and the permeability of the glandular cells. The latter may vary in different physiologic circumstances, and it probably is not the same in the different types of cell.

Secretion is not a simple process of filtration through the capillaries and glandular cells. The following facts are proof of this assertion:

1. The osmotic pressure of the saliva is more or less one-half that of the blood, varying according to the stimulus which has provoked secretion. If the membranes that separate the blood from the glandular duct were semipermeable, to obtain an ultrafiltrate of this osmotic pressure, hydrostatic pressures of ten to twenty times the blood-pressure level would be needed.
2. Some electrolytes are in higher concentration in the secreted fluid than in blood plasma.
3. If the carotid blood pressure and the pressure in the duct of the submaxillary gland are measured simultaneously, it will be seen that on stimulation of the chorda tympani the pressure in the duct rises and can exceed the carotid blood pressure (Ludwig, 1851).
4. Stimulation of the chorda tympani after atropine has been injected does not provoke salivary secretion, although there is intense vasodilatation.
5. Oxygen consumption increases when the glands are secreting. In the so-called "resting period" the cells continue to consume oxygen and synthesize the specific products of their secretion (protein, enzymes, etc.) which are discharged during the secretory period.

Secretion is, therefore, a special type of cellular activity in which there are two main processes:

1. Transference of water and crystalloids from the blood through the walls of the capillaries and glandular epithelium to the lumen of the digestive tract. In some cases products of glandular synthesis, either inorganic (*e.g.*, HCl in the gastric mucosa) or organic (*e.g.*, bile salts in the liver), are also transferred.
2. Release of organic colloids synthesized by the glandular cells (proteins, enzymes, etc.) which are diluted and flushed out by the fluid passing through the cells.

In most of the secretory cells two parts can be distinguished in the cytoplasm: (*a*) a homogenous optically structureless substance surrounding the nucleus, and (*b*) structural elements such as the mitochondria and Golgi apparatus, which form part of the cell, and the granules which are products of cellular synthesis stored until they are discharged in the process of secretion (Fig. 159). During the resting period these granules accumulate around the nucleus. The Golgi apparatus and the mitochondria are thought to have an active part in their formation.

The granules are the specific products of the glandular cells. They are constituted by proteins, some of which are inert proteins and others enzymes, but mostly enzyme precursors (zymo-

gens) which require an activator to become active enzymes.

Innervation. Ludwig[1] observed that stimulation of the chorda tympani, which innervates the submaxillary gland, provokes salivary secretion, thus discovering secretory nerves. Afterward the dual innervation, sympathetic and parasympathetic, of all the digestive glands was demonstrated.

According to Heidenhain[2] the salivary glands are innervated by two kinds of fibers: (*a*) *secretory fibers* which modify cellular permeability and cause water and salts to be transferred from the interstitial periglandular space to the lumen of the digestive tract; and (*b*) *trophic fibers* which modify the metabolism of the cells, causing them to discharge their specific products into the glandular duct. Babkin[3] considers that Heidenhain's terminology may be confusing and suggests it should be replaced by the terms *hydrelatic effects* (ἐλαύνο, to propel) and *ecbolic effects* (ἐκβαλικός, throwing out). The substitution of "effects" for "fibers" has the advantage of including the activity of hormones, chemical mediators, etc. Moreover, the new terms not only are more expressive than those used by Heidenhain but also have a more precise meaning.

The *hydrelatic function* of the digestive glands is regulated by nerves and hormones. It is under nervous control in the large salivary glands and in the gastric glands during the first phase of secretion, and under hormonal control in the pancreas, liver, intestinal glands, and in the gastric glands during the chemical phase of secretion.

The *ecbolic function* is under nervous control in the majority of the glands. At one time the sympathetic was attributed an almost exclusively ecbolic effect, and the parasympathetic hydrelatic effects. Babkin has demonstrated that in nearly all the digestive glands (salivary, gastric, pancreatic, and intestinal) ecbolic effects are controlled by the parasympathetic. Stimulation of the parasympathetic causes the conversion of organic colloids stored in the cells into soluble products which are discharged into the duct. The sympathetic exerts several kinds of effects: (*a*) it has hydrelatic and ecbolic effects on some of the cells of composite glands, *e.g.*, the semilunar cells of the submaxillary glands and the mucous cells of the gastric mucosa; (*b*) in addition to its vasomotor effects, it innervates contractile myoepithelial cells; (*c*) it has cholinergic fibers which release acetylcholine and produce the same effects as the parasympathetic.

[1] Ludwig, C., *Zeit. f. Rat. Med.*, *n.F.*, **1**, 255, 1851.

[2] Heidenhain, R., *Stud. Physiol. Inst. Breslau*, **4**, 1, 1868.

[3] Babkin, B. P., "Secretory Mechanism of the Digestive Glands," 2d ed., Hoeber, New York, 1950.

Motility. The muscles of the digestive tract (both striated and smooth) are disposed in various ways in the different segments. They exert several effects: (*a*) they break up and crush the food; (*b*) they mix food with the digestive secretions; and (*c*) they propel the contents of the digestive tract from the mouth toward the large intestine and finally expel the residue. Striated muscles carry out the function of mastication; they are found in the pharynx and upper part of the esophagus, where they take part in the act of swallowing; finally, the external sphincter of the anus is also a striated muscle. All the other segments are provided with smooth muscles, usually disposed in an external longitudinal layer and an internal circular layer; an oblique layer is occasionally found. The coordinated contractions of these muscles and the movements they produce will be described later.

The different segments are separated by *sphincters*, *i.e.*, thickening of the muscle layers and the mucosa; these are the cardia, pylorus, ileocolic sphincter, and internal and external sphincters of the anus.

Properties of digestive smooth muscle. The smooth muscles of the digestive tract form part of Bozler's visceral smooth muscles, which he considers similar to cardiac muscle, syncytial in structure, and with a well-developed automatism.

When a visceral smooth muscle is stretched, it lengthens rapidly and then more slowly, taking a long time to reach its final length. The tension at first rises but then rapidly falls to a low level. Gradual stretching produces little or no rise in tension. A cavity with smooth-muscle walls, *e.g.*, the stomach, can thus be greatly distended without any significant rise in tension. Sudden stretching can produce a much greater and more lasting increase in tension because it stimulates contraction. There is no constant ratio of tension to length, so a muscle can have the same tension at widely different lengths. The "postural tonus" of a viscus, *e.g.*, the "tonus" of the stomach which determines its shape, is dependent on the length of the fibers, not on the tension developed.

Visceral smooth muscles respond to nerve impulses, direct electric stimulation, sudden

stretching, and chemical substances. Among the latter, chemical mediators of neuromuscular transmission (acetylcholine, adrenaline, noradrenaline) are of special importance. The response, which may be contraction or relaxation according to the stimulus and the condition of the muscle, develops slowly. The latent period lasts more than 100 msec., and contraction lasts several seconds; relaxation is even slower. Usually only a small tension is developed on contraction, even when the muscle shortens considerably. Exceptionally strong and sustained contractions are painful (spasms, colic). The response may be localized, or may spread throughout the muscle, accompanied by a spike potential and followed by a refractory period.

Spontaneous rhythmic contractions are frequently observed in visceral muscles. Their amplitude and frequency varies and may be modified (inhibited or augmented) by nerve impulses and their chemical mediators. They have been observed, however, after all nervous influences have been abolished.

Innervation of digestive muscles. The smooth muscles of the digestive tract are innervated by the sympathetic and parasympathetic (extrinsic innervation). Moreover, there are ganglion cells and nerve fibers in the walls of the digestive tract (intrinsic innervation). There are two nerve plexuses in the intestine: (*a*) Meissner's plexus, situated in the submucosa between the muscularis mucosae and the circular muscle layer; and (*b*) the myenteric or Auerbach's plexus, placed between the circular and longitudinal muscle layers.

These plexuses are formed by (*a*) neurons (ganglion cells), usually considered as part of the parasympathetic; (*b*) preganglionic and postganglionic parasympathetic fibers; (*c*) postganglionic sympathetic fibers from the lateral and prevertebral sympathetic ganglia; and (*d*) afferent fibers.

Strips of muscle fibers from the longitudinal layer, free from nerve cells and nerve fibers, show spontaneous rhythmic contraction. In this case rhythmic activity is undoubtedly myogenic. In the intact organism in normal conditions, however, nerve impulses give rise to, and modify, rhythmic contractions.

Propagated contractions in the organism (peristalsis) require the integrity of the nerve plexuses. Cannon's myoenteric reflex, *i.e.*, contraction of the circular layer above, and relaxation below, a stimulated point, in which the plexus plays a necessary part, may, however, not be a true reflex with an afferent and a motor neuron within the plexus, but an axon reflex in which excitation spreads along nerve fibers.

The extrinsic innervation consists of

1. Preganglionic parasympathetic fibers ending on the neurons of Auerbach's plexus. Stimulation of these fibers increases muscular activity in the digestive tract.
2. Postganglionic sympathetic fibers which end on the muscle fibers. Stimulation may produce contraction or inhibition.

The sympathetic exerts a continuous moderating effect, and its removal is followed by augmented motility for a few days; later the denervated segment behaves normally.

Afferent fibers are small nonmyelinated fibers which enter the spinal cord in the dorsal roots. Ascending paths carry impulses from these fibers up to the cortex, and functional integration occurs at the different levels from the cord to the cortex (see Chaps. 73 and 84). Cortical integration of digestive functions explains how these functions are subject to psychic and emotional influences.

Hormones of the digestive tract. Digestive functions are controlled not only by nerve impulses but also by humoral factors, and most of the digestive functions have a double control, nervous and humoral. Complete denervation is usually followed by only transitory disturbances, which do not affect digestive functions sufficiently to cause serious hindrance to digestion. Nervous control is preponderant in the initial stages of digestion, but it diminishes and humoral factors become more important in the further stages. Salivary secretion is almost exclusively dependent on the nervous system; both nervous and humoral factors exert their effects on the stomach; the pancreas is mainly under the control of humoral influences; in the large intestine, nervous control is again preponderant.

Digestive hormones are among the most important humoral factors that control digestion. Hormones are specific substances secreted by an organ or tissue and carried by the blood to the organ or tissue on which they exert their specific activity. Digestive hormones are produced in the gastrointestinal mucosa in response to specific stimuli.

There are certain differences between digestive hormones and those secreted by the endocrine glands. The endocrine glands constitute an interrelated system; digestive hormones are an independent system which is not influenced by, and has no influence on, the endocrine system. The only known factors which regulate the activity of the digestive hormones are substances in the gastrointestinal tract (food, products of digestion, secretions of the mucosa, etc.). When the adequate stimulus is applied, after a latent period, the hormone is released, and when the stimulus ceases to be active, the hormone ceases to be secreted.

The extraction from a segment of gastrointestinal mucosa of a substance which exerts an effect on some aspect of digestive secretion or motility is not sufficient evidence for attributing hormonal activity to this substance. The following conditions must first be fulfilled:

1. The substance must be released by the effect of specific stimuli, acting on a definite segment of the gastrointestinal mucosa.
2. The substance must exert a specific effect on a certain organ or tissue and not on others.
3. This effect must be produced by the release of the substance from the mucosa into the blood stream.
4. The substance must be extracted and purified sufficiently so as to eliminate other active substances (acetylcholine, histamine, etc.).
5. Experimental evidence must be obtained that in physiologic conditions a certain digestive function is under hormonal control. Several digestive hormones have been described; for some of them, however, conclusive evidence that they play a part in the physiologic control of digestion has not yet been obtained.

The following are the main digestive hormones:

Gastrin (Edkins[1]) is secreted by the gastric mucosa in response to mechanical stimuli (distention of the stomach) and chemical stimuli (products of partial protein digestion, meat extract, etc.); it stimulates the secretion of HCl by the gastric glands.

Secretin (Bayliss and Starling[2]) is released by the mucosa of the upper part of the small intestine when stimulated by certain chemical agents (HCl, etc.); it provokes the secretion of pancreatic juice.

Pancreozymine (Harper and Raper[1]) is released by the intestinal mucosa stimulated by chemical agents (soaps, peptone, starch, etc.); it stimulates the secretion of pancreatic enzymes.

Cholecystokinin (Ivy and Oldberg[2]) is released by the mucosa of the upper part of the small intestine stimulated by certain chemical agents (fats, fatty acids, HCl, peptone, etc.); it provokes contraction and emptying of the gall bladder.

Enterogastrone (Kosaka and Lim[3]) is released by the mucosa of the upper part of the small intestine stimulated by fat and sugars; it inhibits gastric secretion and motility.

Enterocrinin (Nasset[4]) is released by the intestinal mucosa in response to the ingestion of a mixed meal; it increases the volume and concentration of intestinal juice.

Villikinin (Kokas and Ludany[5]) is released by the duodenal mucosa in response to HCl; it increases motility of intestinal villi.

Duocrinin (Grossman[6]) is released by the intestinal mucosa stimulated by certain foods (mixed meal, dextrose, cottonseed oil, etc.); it stimulates secretion of the Brünner glands in the duodenum.

It is doubtful that other substances can be considered as hormones, *e.g.*, *urogastrone*, extracted from urine, which depresses gastric motility and secretion, and *anthelone*, extracted from urine and intestinal mucosa, which has antiulcer activity in dogs with a Mann-Williamson operation.

ABSORPTION

Absorption consists in the passage into the organism of substances coming from the environment. The skin does not absorb water or substances dissolved in water. Gases and volatile substances (iodine, methyl salicylate, etc.) and fat-soluble substances (Hg, steroids) are absorbed through the skin, a fact that has important applications in therapeutics. The linings of the body cavities (pleura, peritoneum, etc.)

[1] Edkins, J. S., *Proc. Roy. Soc., London, s.B.*, **76**, 376, 1905.

[2] Bayliss, W. M., and E. H. Starling, *J. Physiol.*, **28**, 325, 1902.

[1] Harper, A. A., and H. S. Raper, *J. Physiol.*, **102**, 115, 1943.

[2] Ivy, C., and E. Oldberg, *Proc. Soc. Exper. Biol. & Med.*, **25**, 251, 1928.

[3] Kosaka, T., and R. K. S. Lim, *Proc. Soc. Exper. Biol. & Med.*, **27**, 890, 1930.

[4] Nasset, E. S., *Am. J. Physiol.*, **121**, 481, 1938.

[5] Kokas, E., and G. Ludany, *Compt. rend. Soc. de biol.*, **232**, 293, 1932.

[6] Grossman, M. J., *Physiol. Rev.*, **30**, 33, 1950.

and mucosae (conjunctiva, respiratory tract, sublingual mucosa, etc.) rapidly absorb substances dissolved in water, oil, or glycols.

Absorption from the digestive tract. The purpose of digestion is to convert ingested foodstuffs by a long series of chemical reactions into simple, easily soluble substances that can be readily absorbed. The passage of these substances into the cells lining the digestive tract and their transfer to the internal media are accomplished by physicochemical processes that are still not completely understood. The problem of digestive absorption is part of the problem of cellular permeability which is as yet unsolved.

The stomach can absorb fairly rapidly alcohol and other substances[1] such as alkaloids, salts, acids, and the products of protein digestion; it absorbs little or no sugar, fat, or carbohydrate. The stomach is of only secondary importance in the absorption of foodstuffs.

The large intestine can absorb water and certain substances in solution such as salts, glucose, etc.

The greater part of digestive absorption takes place in the small intestine. It is therefore important to study the structure of the large mucous surface of this part of the digestive tract.

The intestinal villi facilitate absorption considerably. Each square millimeter of intestine represents an absorbing surface of 3 to 18 sq. mm. The villi are made up of a supporting frame of connective tissue which contains a few muscle fibers. A small arteriole penetrates into each villus and ends in a capillary plexus drained by one or two venules. There is also a capillary lymph vessel (lacteal). The lacteal vessel opens into a lymphatic network provided with a system of valves that prevents the lymph from returning to the villi. The fingerlike process is covered by a layer of cylindrical epithelial cells with a striated border, lying on a basal membrane; these are the absorbing cells. A few cuplike cells that secrete mucus are interspersed in the columnar epithelium. Between the bases of the villi, the crypts or Lieberkühn's glands open into the intestine. There are four types of cell in these glands: cylindrical epithelial cells, caliciform cells (which secrete mucus), the cells of Paneth, and granular argentophil cells.

Fibers from the muscularis mucosa are inserted on the wall of the lymph vessel of the villi and on the basal membrane. Contraction of these fibers empties the lymph vessels, and their relaxation produces a negative pressure within these vessels which helps absorption from the intestine.[1]

The intestinal villi have a pistonlike movement; they shorten without any considerable thickening, and then gradually regain their length. The effect of these movements on absorption is not well known. The contraction of the smooth-muscle fibers in the villi is responsible for their movement. The fibers respond to stimulation of Meissner's plexus, and it seems that there is a hormone, known as "villikinin,"[2] secreted by the intestinal mucosa, which is the normal stimulus for the contraction of the villi.

Absorption of water and watery solutions. Water and salts are absorbed by the cells covering the villi and apparently pass directly into the blood capillaries and thence to the portal vein. It is possible, however, that these substances pass first into the lymph vessel and that in the submucosa, where lymph vessels and veins are in close contact, an exchange between lymph and plasma can take place. At any rate, water and salts must be taken up mainly by the blood stream, for the ingestion of large amounts of water can dilute the blood, but the absorption of saline solutions does not increase the lymph flow in the thoracic duct.

An important question is whether the process of absorption through the epithelial cells of the intestinal mucosa is exclusively a physical one of osmosis, diffusion, adsorption, imbibition, filtration, etc., or a physiologic process in which the cells take an active part. This active participation of the cells might consist in the selective absorption of those substances most urgently needed by the organism, or it might consist simply in the motor mechanism described above.

In the small intestine osmosis and diffusion take place. When a concentrated NaCl solution is injected into an intestinal loop previously isolated from the rest of the bowel, the loop fills with fluid because of the high osmotic pressure of its contents, and NaCl passes into the blood according to the laws of diffusion. Nevertheless the difference in concentration of the intes-

[1] KAREL, C., *Physiol. Rev.*, **28**, 433, 1948.

[1] VERZÁR, F., "Absorption from the Intestine," Longmans, New York, 1936.

[2] KOKAS, E., and G. LUDANY, *Pflüger's Arch. f. d. ges. Physiol.*, **234**, 182, 1934; *Compt. rend. Soc. de biol.*, **122**, 413, 1936.

tinal contents on one side, and the blood or lymph on the other, is not enough to explain absorption in all cases. Heidenhain[1] injected a dog's own plasma into an isolated loop of its intestine and observed that the plasma was completely reabsorbed; in this case there were no differences in concentration to explain absorption; therefore, osmosis is not the main factor in the absorption of water and electrolytes, and other active processes must take place in the intestinal epithelium. More recent experiments, in which radioactive isotopes and heavy water have been used, have confirmed this conclusion.[2] If isotonic solutions of NaCl and Na_2SO_4 are injected into the isolated loop, the NaCl is completely absorbed while the sulfate remains in the bowel. According to the laws of diffusion it would be expected that chloride, of which there is 0.6 per cent in the blood, would be absorbed slowly, while the sulfate, of which there is very little in the blood, would be rapidly absorbed.

The velocity of absorption by the intestine of many substances (*e.g.*, glucose) remains constant throughout the whole process; if absorption took place by diffusion and osmosis, the velocity would decrease as the concentration of the substance in the intestine diminished. When the cells of the mucosa are intoxicated with cyanide or fluoride or there is intense anemia, the mucosa behaves as an organic but lifeless membrane. Therefore in the passage of water and salts from the intestine to the blood, processes dependent on the integrity of the metabolism of the cell play a part. What these processes consist in, and what forces take part in them, are problems still unsolved; probably they are similar to those by means of which glomerular filtrate is reabsorbed by the renal tubes.

Evidence has been given of the selective absorption of several substances. For example, glucose and galactose are absorbed more rapidly than other carbohydrates. Phosphorylation, taking place in the cells of the mucosa, has been considered as the cause of this selective absorption. Phosphorylation consists in the union of the carbohydrate and phosphate to form hexosephosphates. Monoiodacetic acid inhibits phosphorylation and suppresses the selective absorption of monosaccharides.

[1] Heidenhain, R., *Pflüger's Arch. f. d. ges. Physiol.*, p. 579, 1894.

[2] Visscher, M. B., *et al.*, *Am. J. Physiol.*, **142**, 550, 1944; **144**, 457, 1945.

Another problem still under discussion is whether the absorbed substances pass from the lumen of the intestine into the blood through the cells or between them. The epithelial cells, like other cells in the organism, are covered by a membrane that is permeable only to fat-soluble substances. Therefore water, salts, carbohydrates, and other substances not soluble in fats would have to pass between the cells. Experiments with dyes, soluble and insoluble in fats, seem to confirm this assertion. It is also possible that the lipids in the external layer of the epithelium act as a protective mechanism and that the properties of the membrane are modified by the needs of the organism.

Absorption of proteins.[1] The products of the digestion of protein are absorbed exclusively by the small intestine.

Observations made on animals with fistulas at different levels of the small intestine show that ingested protein is completely or almost completely changed by the action of the digestive enzymes into amino acids or very simple peptides.

During the period of absorption there is a slight and transitory increase in the amino-acid concentration of the blood. The lymph vessels cannot be of much importance in the absorption of protein, as they can be totally suppressed without causing serious interference in protein absorption.

The fact that an organism can be maintained in protein equilibrium by intravenous injection of amino acids, or of the products obtained by digesting proteins *in vitro*, gives further support to the contention that proteins are absorbed as amino acids. This procedure is used in the feeding of patients whom it is not possible to feed by mouth.

Nevertheless, immunologic methods that allow the detection of very small amounts of specific proteins have been used to show that normally in many subjects certain proteins can pass unaltered from the intestine into the blood. This fact is of importance, inasmuch as it proves the possibility of sensitization to ingested proteins (alimentary allergy), but it does not modify substantially the assertion that proteins are absorbed after they have been disintegrated into their component amino acids.

Absorption of fats.[2] Neutral fats in the food are split up into their components—glycerol and

[1] See Chap. 43, Protein Metabolism.

[2] See Chap. 42, Fat Metabolism.

fatty acids—by the action of the digestive lipases, especially of pancreatic lipase. Examination of the intestinal contents at different levels has shown that, as the distance from the stomach increases, the proportion of fatty acids to fats increases. The fatty acids form, with the bile salts, diffusible and water-soluble sodium soaps. These soaps facilitate the emulsification of fats, which therefore become more susceptible to the action of enzymes. Soaps are unstable because of the slight acidity of the intestinal contents. Monoglycerides also facilitate fine emulsification of fats.

Glycerol and fatty acids are taken up by the epithelial cells, where they again combine to form fats, which pass into the lymphatics. After a meal containing fats the lymphatics in the mesentery are filled with a milklike fluid—the chyle—the opacity of which is due to small particles of 0.05 to 0.1 μ, called chylomicrons. The milky fluid is an emulsion of neutral fats. After a fatty meal, lymph loaded with fat flows in increased amount from the thoracic duct, and the fat content of the blood plasma is increased. The mechanism of the absorption of fats will be discussed when considering fat metabolism (Chap. 42).

HUNGER AND APPETITE

Cannon and Washburn[1] demonstrated that the painful sensations of hunger coincide with strong peristalsis in the body of the stomach. Carlson and his associates investigated factors that awake or inhibit the painful contractions. The theory thus arose that hunger is essentially a painful gastric sensation. The behavior of men and animals after gastrectomy or denervation of the stomach, the fact that insulin increases the ingestion of food in animals with a denervated stomach whereas sympathicomimetic amines decrease it, and other observations show that hunger is something more than painful gastric contractions.

Body weight is maintained constant during long periods in adult life, in spite of considerable variations in the consumption of energy; therefore the ingestion of food is adjusted with great accuracy to the needs of the organism. "The hunger state is the physiological state resulting from the privation of food of a specific or general type and abolished by the ingestion of these foods" (Ivy). This is a primitive, nonconditioned state, made evident by (*a*) special behavior, *e.g.*, increased motility; and (*b*) hunger sensations, such as general weakness, tiredness, nausea, irritability, headache, sensation of emptiness or tension, and pain in the epigastrium.[1]

[1] Cannon, W. B., and A. L. Washburn, *Am. J. Physiol.*, **29**, 441, 1912.

After experience has taught man and animals that ingestion of food suppresses the disagreeable sensations and behavior of hunger, conditioned cortical reactions increase the complexity of the response. Hunger sensations give rise to the desire for food or *appetite*, and hunger behavior is incorporated into an acquired *appetitive behavior* (Janowitz and Grossman). The ingestion of food suppresses appetitive behavior, and the desire for food ceases (satiety or physiologic anorexia).

The ingestion of food is normally controlled so that a stable equilibrium is maintained between the intake and output of energy, and the body weight remains fairly constant in spite of wide variations in the diet, the work performed, and the temperature of the environment. There are regulatory mechanisms in the central nervous system which control this adjustment.[2] In dogs and other animals the destruction of part of the ventromedial nucleus of the hypothalamus provokes hyperphagia and obesity.[3] Destruction of the lateral aspect of the lateral region of the hypothalamus produces complete anorexia in cats and rats.[4] This part of the hypothalamus acts as a feeding center, normally inhibited by those parts of the ventromedial nucleus the destruction of which causes hyperphagia. Evidence has been brought forward of the existence in the hypothalamus of chemoreceptors sensitive to changes in the blood-sugar level, and to other humoral factors.[5]

Specific appetite.[6] The constancy of the internal environment (homeostasis) is maintained by a series of physiologic and chemical

[1] Janowitz, H., and M. I. Grossman, Hunger and Appetite: Some Definitions and Concepts.

[2] Brobeck, J. R., *Physiol. Rev.*, **26**, 541, 1946.

[3] Camus, J. R., and G. Roussy, *Comp. rend. Soc. de biol.*, **75**, 483, 1913; Hetherington, A. W., and S. W. Ranson, *Proc. Soc. Exper. Biol. & Med.*, **41**, 465, 1939.

[4] Anand, B. K., and J. R. Brobeck, *Yale J. Biol. & Med.*, **24**, 123, 1951.

[5] Mayer, J., J. J. Vitale, and M. W. Bates, *Nature, London*, **167**, 562, 1951.

[6] Richter, C. P., Total Regulatory Functions in Animals and Human Beings, *Harvey Lect.*, **38**, 63, 1942–1943.

regulatory mechanisms. Numerous examples of coordinated processes that regulate the blood pressure, the blood sugar, the acid-base equilibrium, etc., can be brought forward; besides these, the organism as a whole presents reactions that contribute to maintain homeostasis. These are reactions of behavior that have been studied systematically by Richter, corroborating and extending previous isolated observations.

The existence of this type of regulation, in which the whole organism takes part, was demonstrated by Richter in animals after a disturbance in the equilibriums of their internal environment had been induced by the extirpation of an organ of fundamental importance for homeostasis. For example, removal of the adrenal glands in rats produces a disturbance in sodium metabolism; the animal loses great quantities of sodium in the urine and the internal environment is seriously upset. If the animals are given water and an ordinary diet, they die in 8 to 15 days. If on the other hand they are also given access to a 3 per cent solution of NaCl, they will drink enough of this saline solution to keep alive and without symptoms of adrenal insufficiency.

Another example is given by parathyroidectomized rats, in which there is a serious disturbance in calcium metabolism. These animals are put into a cage where there are two graduated drinking flasks, one with distilled water and the other with a 2.4 per cent calcium lactate solution, and the amount of fluid drunk from each flask is measured daily. Immediately after parathyroidectomy, the animals drink more calcium lactate and thus keep alive and free from symptoms of tetany. If parathyroid insufficiency is now treated in these animals by the administration of a drug (dihydrotachisterol), or by grafting a parathyroid in the anterior chamber of the eye, the animals will reduce the amount of calcium lactate they drink to that taken before parathyroidectomy (Fig. 160).

FIG. 160. Specific appetite for calcium lactate. Parathyroidectomy increases the ingestion of calcium, and parathyroid implantation decreases it. (*Richter, C. P., Harvey Lect., vol. 38, p. 63, 1942–1943.*)

The capacity of rats to select their diet goes even further. Animals were put in cages with several drinking flasks and feeding troughs, in which measured amounts of protein, carbohydrate, fat, vitamins, and several salt solutions were placed. The rats chose for themselves a complete diet which allowed them to grow at the same rate as the controls fed on a standard diet. Pancreatectomized rats spontaneously re-

fused carbohydrate and fed principally on protein and fat. When vitamin B_1 (thiamine, aneurine) was withheld, the animals diminished the ingestion of carbohydrate and increased that of fat.

These examples are enough to show the practical importance of a method that can be used in many ways and that has already been admirably applied by Richter and his collaborators. The mechanism of these self-regulatory functions is still incompletely understood.

BIBLIOGRAPHY

ALVAREZ, W. C., "An Introduction to Gastroenterology," 4th ed., Hoeber, New York, 1948.

BABKIN, B. P., "Secretory Mechanism of the Digestive Glands," 2d ed., Hoeber, New York, 1950.

CANNON, W. B., "Hunger and Thirst," Clark University Press, Worcester, Mass., 1934.

CANNON, W. B., "The Mechanical Factors of Digestion," E. Arnold & Co., London, 1911.

CARLSON, A. J., "The Control of Hunger in Health and Disease," Chicago University Press, Chicago, 1916.

FISHER, E., Vertebrate Smooth Muscle, *Physiol. Rev.*, **24**, 467, 1944.

GOLDSCHMIDT, S., On the Mechanism of Absorption from the Intestine, *Physiol. Rev.*, **1**, 421, 1921.

GREENGARD, H., The Hormones of the Intestinal Tract, in "The Chemistry and Physiology of Hormones," American Association for the Advancement of Science, Washington, D. C., 1944.

GROSSMAN, M. I., Gastrointestinal Hormones, *Physiol. Rev.*, **30**, 33, 1950.

HOUSSAY, B. A., Hormonas del Aparato Digestivo, *Prensa méd. argent.*, **28**, 1074, 1944.

IVY, A. C., The Role of Hormones in Digestion, *Physiol. Rev.*, **10**, 282, 1930.

PAVLOV, I. P., Die äussere Arbeit des Verdauungsdrüsen und ihr Mechanismus, in Nagel's "Handbuch der Physiologie des Menschens," **2**, 666, Friedrich Vieweg und Sohn, Brunswick, 1907.

PAVLOV, I. P., "The Work of the Digestive Glands," 2d ed., Griffin, London, 1910.

RICHTER, C. P., Total Self-regulatory Function in Animals and Human Beings, *Harvey Lect.*, **38**, 63, 1942–1943.

RYLE, C. P., "Gastric Function in Health and Disease," Milford, London, 1926.

VERZÁR, F., "Absorption from the Intestine," Longmans, New York, 1936.

CHAPTER 36

The Secretion of Saliva. Mastication and Deglutition

THE PROCESS OF digestion begins in the mouth where food is soaked in saliva, is broken up and crushed by mastication, and after the bolus has been formed, is swallowed, *i.e.*, passed into the pharynx and through the esophagus to the stomach (deglutition).

THE SECRETION OF SALIVA

Saliva is secreted by the three pairs of salivary glands, the parotid, submaxillary, and sublingual glands, the names of which indicate approximately their position. There are also small "buccal" glands situated in the mucosa of the mouth. The principal salivary glands pour their secretions into the mouth by means of well-differentiated ducts, which can be catheterized so as to collect the secretions. The ducts of the submaxillary and sublingual glands end in the floor of the mouth, under the tongue; the parotid ducts end on the side wall of the mouth on each side.

Methods of study. There are two ways in which the saliva can be studied: (*a*) the total, mixed saliva can be collected by voluntary or forced evacuation of the mouth (spitting) or by means of a fistula of the esophagus; (*b*) the saliva secreted by each gland can be obtained by catheterization of the respective duct or from a fistula produced experimentally[1] (Fig. 161) or pathologically.

Types of gland. Three types of gland can be distinguished according to their structure: mucous, serous, and mixed.

In the mucous glands the cells lining the walls of the alveoli form a single layer on the basal membrane, and the cytoplasm contains numerous droplets of mucinogen. The serous glands have smaller cells, with a deeply staining nucleus and a granular cytoplasm (zymogen granules). In man the buccal glands are of both types; the parotid is a typical serous gland; the submaxillary and sublingual glands are mixed, but serous cells predominate in the former and mucous cells in the latter.

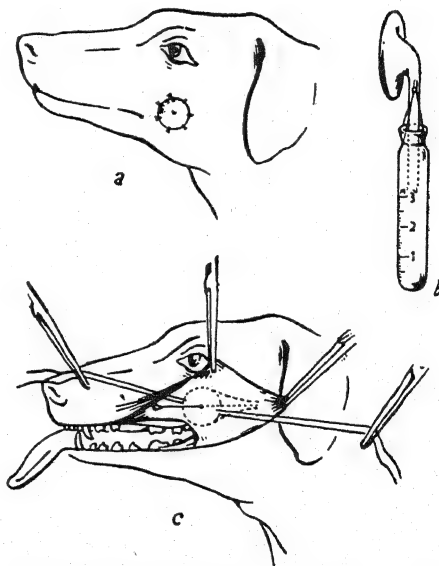



FIG. 161. Parotid fistula in the dog. (*Le Play, "Technique opératoire physiologique," Masson et Cie, Paris, 1912.*)

Types of saliva. The secretion of the mucous glands is viscous, because of its high mucin content. In man it also contains an amylase, ptyalin.

[1] LE PLAY, "Technique opératoire physiologique," Masson et Cie, Paris, 1912, pp. 72 and 79.

Serous glands, *e.g.*, the parotid, secrete a fluid saliva, with almost no mucin; on the other hand the ptyalin content of the parotid secretion is about four times that of the mucous glands.

Stimulation of salivary secretion. In man saliva is secreted continuously (saliva of fasting or rest); the amount secreted is approximately 15 cc. per hr. With adequate stimuli this amount can be rapidly and considerably increased. The quantity and quality of the saliva vary with the nature of the stimulus. The amount secreted in 24 hr. ranges from 600 to 1,500 cc. During sleep almost no saliva is secreted. An increase in salivary secretion is obtained by applying thermal (heat), mechanical (chewing sand, stones, paraffin, rubber, etc.), or chemical stimuli to the mucosa of the mouth. Among the chemical stimuli the most efficient are those provoking gustatory sensations. Substances not actually foods, such as acids, alkalis, bitters, etc., which have a strong disagreeable taste cause abundant secretion of saliva. Dryness of the mouth or irritation of the mucosa by defective teeth or badly placed dental fixtures can also provoke abundant salivary secretion. Distention of the esophagus and stimulation of the gastric mucosa by food or irritating substances stimulate the secretion of saliva. The flow of saliva increases whenever there is an increase in the acidity of the blood.[1]

Conditioned reflexes produce salivary secretion, as when the "mouth waters" at the sight or thought of food. In an animal with a salivary fistula, the flow of saliva produced by the introduction into the mouth of a savory substance or of food can be measured. This secretion is the result of an inborn reflex, which has its center in the medulla and takes place without the intervention of the cerebral cortex. If the introduction of a salivary stimulus into the mouth is repeatedly associated with the application of another stimulus, incapable of producing salivary secretion per se, *e.g.*, the ringing of a bell, after several repetitions salivary secretion will be provoked by just ringing the bell, even if no food is given. The animal thus acquires a new reflex, called by Pavlov a "conditioned reflex." A stimulus originally unefficacious takes on by association the properties of the stimulus that naturally produces salivary secretion. Such a "conditioned reflex" occurs only if the corresponding parts of the cerebral cortex remain intact.

[1] Brassfield, C. R., *Am. J. Physiol.*, **144**, 43, 1945.

Reflex nature of salivary secretion. The stimuli mentioned above produce salivary flow by reflex action. The afferent paths are the sensory nerves of the mucosa of the mouth. The glossopharyngeal nerve innervates the posterior third of the tongue; the anterior two-thirds and the tip of the tongue are innervated by the lingual nerve (the gustatory fibers being included in the chorda tympani).

The salivary center is situated in the reticular formation of the fourth ventricle, between the nucleus of the facial nerve and Deiter's nucleus. The cephalic part of this nucleus is connected with the submaxillary glands, and the caudal part with the parotid glands.[1] Salivary flow can be produced by stimulation of the cerebral cortex, but section of the brain stem above the pons does not suppress the normal inborn salivary reflexes; therefore the most important center is the one situated in the medulla and pons.

The salivary glands are innervated by the visceral nervous system, each gland receiving both sympathetic and parasympathetic fibers (Fig. 162).

The parasympathetic fibers of the submaxillary and sublingual glands are conveyed by the chorda tympani; they come from the salivary center in the pons, branch off from the facial nerve with the chorda tympani, and eventually join the lingual nerve. In the floor of the mouth the preganglionic parasympathetic fibers branch off from the lingual nerve going toward the submaxillary gland and ending in a small ganglion situated in the hilum of the submaxillary gland. Fine unmyelinated postganglionic secretory fibers go from this ganglion and end in the gland. The postganglionic fibers for the sublingual gland also arise in the submaxillary ganglion.

The parasympathetic fibers for the parotid have their origin in the medulla and are incorporated into the glossopharyngeal nerve, passing to its tympanic branch or nerve of Jacobson, and thence through the small superficial petrosal nerve to the otic ganglion. The postganglionic fibers commencing in this ganglion join the auriculotemporal nerve, a branch of the inferior maxillary division of the fifth nerve.

The sympathetic fibers for the three salivary glands leave the spinal cord in the second to the

[1] Wang, S. C., *J. Neurophysiol.*, **6**, 195, 1943.

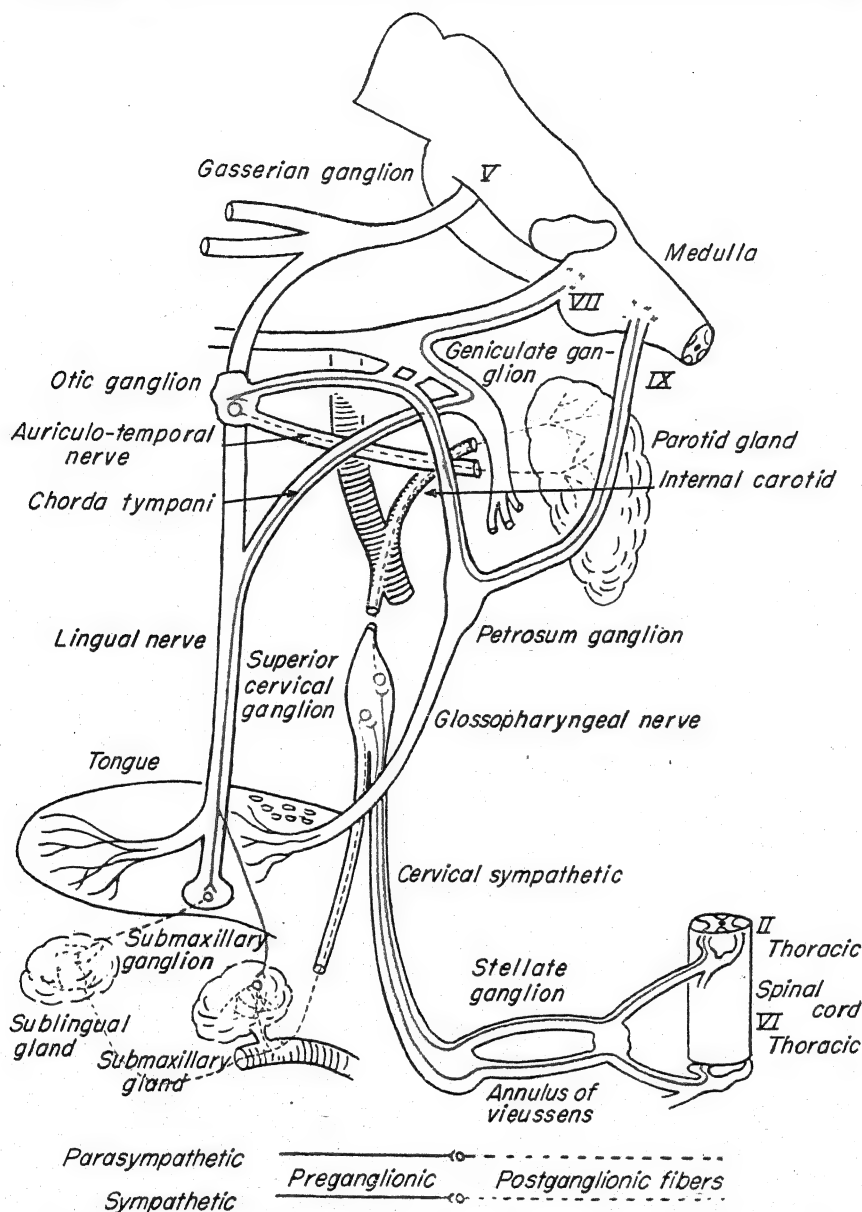



Fig. 162. Innervation of the salivary glands. (*Redrawn from Müller, with modifications.*)

sixth dorsal roots. They end at the superior cervical ganglion, whence the postganglionic fibers follow the branches of the external carotid to the respective salivary gland.

The parasympathetic nerves are vasodilator and the sympathetic nerves are vasoconstrictor.

Electrical stimulation of the secretory nerves. Electrical stimulation of the chorda tympani provokes in a few seconds abundant salivary secretion by the submaxillary gland. The concentration of the saliva and the quantity secreted depend to some extent on the intensity of the current used for stimulation. An increase in the strength of the stimulus provokes a more abundant secretion, not only of fluid but also of the organic and inorganic components of saliva. The vessels of the gland dilate, the blood flow increases from four to eight times, and the oxygen consumption increases from two to three times.

Stimulation of the auriculotemporal nerve produces similar effects in the parotid.

Stimulation of the sympathetic nerves produces the secretion by the submaxillary and sublingual glands of a small quantity of viscous saliva, but no secretion in the parotid gland. A marked vasoconstriction is provoked in all three glands.

The action of the parasympathetic is due to the liberation of acetylcholine in the nerve endings; injection of acetylcholine produces exactly the same effect as stimulation of the parasympathetic nerve. The effects of acetylcholine and stimulation of the parasympathetic secretory nerves are reinforced by eserine and suppressed by atropine. Therefore the parasympathetic fibers of the salivary glands are cholinergic. Pilocarpine also provokes profuse salivation, but the saliva secreted is not identical to that secreted as a result of acetylcholine injection or stimulation of the parasympathetic.

The effect of sympathetic stimulation is due to the release of a chemical mediator. This is demonstrated by the following experiment: If the cervical sympathetic is stimulated after cutting all its branches except those to the submaxillary gland, a substance appears in the blood which provokes contraction of the nictitating membrane. This effect is potentiated by cocaine.

The nature of salivary secretion. *Phenomena accompanying glandular activity.* We have seen that stimulation of the parasympathetic produces an intense vasodilatation of the salivary glands and that the blood flow becomes four to eight times greater. The increase in the blood flow occurs even if the secretory action is inhibited by atropine. The volume of the gland increases, and an electrical variation can be registered. The oxygen consumption is increased; Barcroft found that from 0.25 cc. per min. it rose to 0.86 cc. per min. in the submaxillary gland of a small dog.

There are also *histologic changes* during activity. In the resting gland the cytoplasm is full of zymogen or mucinogen granules. As the gland becomes active the granules diminish until they are reduced to a few in the pole of the cell nearest the duct. These granules are usually considered as made up of the precursors of ptyalin in the serous cells and mucin in the mucous cells.

As far back as 1851, Ludwig[1] saw that the electrical stimulation of the chorda tympani increased salivary flow, and his experiments demonstrated that this effect was not due merely to filtration of water and salts from the blood through the epithelium of the glands.

Heidenhain was able to distinguish two main processes: (*a*) the passage of water and crystalloids from the blood through the capillary wall and glandular cells into the ducts of the gland, and (*b*) the discharge of organic substances secreted and stored in the gland during the periods of rest. The mechanisms of these processes are probably complex. Nerve impulses and humoral factors may give rise to osmotic forces within the cells, or changes in permeability, which cause the passage of water and crystalloids into the duct. According to Heidenhain, nerve impulses produce "trophic" effects which convert organic colloids stored in the gland into more soluble substances, which are then released into the ducts.

The salivary glandular cells undoubtedly consume energy not only for the synthesis of their secretory products but also for osmotic and mechanic work, because stimulation of their nerves, with the release of chemical mediators, increases the activity of the cells. Acetylcholine increases the O_2 consumption of surviving slices of salivary-gland tissue *in vitro*,[1] a fact which has been interpreted as evidence that the chemical mediator accelerates intracellular metabolic processes. The sources of energy in the salivary gland are apparently similar to those of muscle.[2]

Regulation of salivary secretion. Heidenhain's experiments on electric stimulation of the salivary nerves led him to maintain there were two types of nerve fibers, secretory and trophic. Secretory fibers predominate in the parasympathetic and trophic fibers in the sympathetic, and each cell is innervated by both kinds of fiber. The vasomotor effects exerted by the nerves are relatively independent of their simultaneous secretory or trophic activity.

This theory was almost universally accepted, in spite of the objections brought against it by Langley and Carlson and their associates. Babkin has shown that the trophic (ecbolic) effect is due not to the sympathetic but to the parasympathetic (see page 327). The secretion of a salivary gland in which the sympathetic innervation has been eliminated varies in volume and

[1] Ludwig, C., *Zeit. f. Rat. Med., n.F.*, **1**, 255, 1851.

[1] Deutch, W., and H. S. Raper, *J. Physiol.*, **92**, 439, 1938.

[2] Northrup, D., *Am. J. Physiol.*, **114**, 46, 1935.

amount of organic substances according to the strength of the stimulus applied to the remaining parasympathetic fibers, or according to the stimulus (food, sapid substances) applied in the mouth. These and other results obtained by Baxter[1] and other associates of Babkin led the latter to a new conception of how salivary secretion is regulated, which may be extended to all the digestive glands. A gland does not function as a single entity but is made up of different epithelial components which form distinct functional units. Each one of these units contributes to the formation of the final product in different ways varying with the nerve or humoral stimulation to which they have been submitted.

Histophysiological studies have brought forth evidence that the sympathetic and parasympathetic innervate different groups of cells. Stormont[2] has shown that in the submaxillary gland of the rabbit the chorda tympani innervates the blood vessels, the intercalated ducts, and the homeochrome serous cells. The sympathetic innervates the blood vessels and the tropochrome serous cells. Rawlinson[3] observed the histologic changes provoked by nerve stimulation in the submaxillary gland of the cat. Stimulation of the parasympathetic (chorda tympani) caused changes in the mucous cells, and sympathetic stimulation (cervical sympathetic or intravenous injection of adrenaline) caused changes in the serous (semilunar) cells.

Paralytic secretion. Claude Bernard[4] observed that 2 to 3 days after the chorda tympani had been cut the submaxillary gland began to secrete continuously small amounts of fluid, turbid saliva, which increased gradually, reaching a maximum in 7 to 10 days, and persisted for several weeks. The glands did not respond to reflex stimulation and diminished progressively in size. Section of the sympathetic was not followed by "paralytic secretion." Bernard interpreted these facts as evidence of an inhibitory effect of the nerves on salivary secretion. Heidenhain maintained that saliva retained in the glands after denervation stimulated the cells. Langley saw that stimulation of the peripheral end of the chorda tympani, several weeks after

[1] BAXTER, H., *Am. J. Physiol.*, **91**, 132, 1929; **97**, 450 and 668, 1931.

[2] STORMONT, D. L., The Salivary Glands, in Cowdry's "Special Cytology," 2d ed., Hoeber, New York.

[3] RAWLINSON, H. E., *Anat. Rec.*, **57**, 289, 1933.

[4] BERNARD, C., *Compt. rend. Acad. d. Sc.*, **55**, 341, 1862.

it had been cut, increased the paralytic secretion of the gland. He attributed this to the existence of ganglionic neurons in the nerve and in the gland which were permanently active.

Histological studies[1] of the submaxillary gland after section of the chorda tympani have shown that the mucous cells have the aspect typical of the resting condition and that the serous (semilunar) cells appear to be active. Rawlinson, as was mentioned above, also saw that sympathetic stimulation and intravenous injection of adrenaline increase the activity of the semilunar cells. According to Babkin, paralytic secretion is further evidence that the sympathetic and parasympathetic innervate different groups of cells. Paralytic secretion would, therefore, be due not to the denervated cells but to others which are innervated by fibers not included in the chorda tympani.

Emmelin[2] attributes paralytic secretion, not to spontaneous activity of the cells or their innervation, but to the action of adrenaline or noradrenaline on secretory cells sensitized to these substances by section of the chorda tympani. Babkin and his associates had already observed that parasympathetic denervation sensitized the gland to adrenaline and to sympathetic stimulation. According to Emmelin paralytic secretion is not continuous, but provoked by experimental procedures which cause the discharge of impulses in the sympathetic fibers of the gland, or increase the concentration of adrenaline or noradrenaline in the blood.

Augmented secretion. Stimulation of the chorda tympani increases the response to subsequent stimulation of the same nerve or of the sympathetic.[3] This "augmented secretion"[4] is also observed when sympathetic stimulation precedes sympathetic or parasympathetic stimulation. After stimulation of the chorda tympani possibly the chemical mediator (acetylcholine) persists for a time and adds its effect to that of acetylcholine released by a second period of stimulation. There is also evidence that previous activity of the mucous cells increases the excitability of the semilunar cells, a fact that would explain augmented secretion in response to

[1] MAXIMOW, A., *Arch. f. mikr., Anat.*, **58**, 1, 1901; RAWLINSON, H. E., *J. Anat.*, **70**, 143, 1935.

[2] EMMELIN, H., *Physiol. Rev.*, **32**, 21, 1952.

[3] BRADFORD, J. R., *J. Physiol.*, **9**, 287, 1888.

[4] LANGLEY, J. N., *J. Physiol.*, **10**, 291, 1889.

sympathetic stimulation following stimulation of the chorda tympani.

Changes in saliva in relation to the intensity of the stimulus. On increasing the strength of the electrical stimulus, the amount of saliva secreted and the salt concentration increase. At first the concentration of protein also increases, but later, when the gland is fatigued, it diminishes. Potassium concentration in saliva varies independently of the rate of secretion.

The saliva secreted by the submaxillary gland due to sympathetic stimulation has a higher concentration in K, Ca, mucin, and albumin, and a lower concentration in Na and Cl, than parasympathetic saliva obtained by stimulation of the chorda tympani. These nerves must be stimulated repeatedly to obtain an abundant flow; there seems to be an optimum frequency, which for the chorda tympani of the cat is about nine stimuli per second.

Changes in saliva in relation to the nature of the stimulus. Pavlov called attention to changes in quality and quantity of saliva according to the nature of the substance introduced into the mouth and acting as a stimulus. Raw meat produces a less abundant flow than meat powder. Smooth stones do not provoke salivary secretion, while sand is a potent stimulus. In the case of meat powder abundant salivation facilitates the rapid formation of the bolus, while in the case of sand its removal from the mouth is hastened. Acids provoke the secretion of saliva rich in proteins that act as buffers, so that the excessive acidity tends to be neutralized. Edible substances, such as meat and milk, stimulate the secretion of saliva with a high concentration of organic substances (ptyalin, mucin, etc.) especially adapted to lubricate the food (lubricant saliva). Dry substances (meat powder, sand) produce an abundant flow of fluid saliva (diluting or solvent action).

The mechanism by which this fine quantitative and qualitative adaptation to the nature of the stimulus takes place is as yet unknown.

FUNCTIONS OF SALIVA

Saliva has many important functions, but is not essential to life in man; cases have been described of congenital absence of the salivary gland (xerostomia or aptyalism) in which there was no serious disturbance in digestion. The dog and the rabbit survive extirpation of the salivary glands and remain apparently normal. On the other hand if both parotid ducts are fistulized in a horse, the animal has difficulty in chewing and swallowing, it eats less and more slowly and eventually can become seriously ill. Removal of the salivary glands in the newborn rat disturbs swallowing to such an extent that the animal dies of starvation. The suppression of the flow from both parotids in the sheep produces after a time a progressive state of undernourishment which ends with the death of the animal. This has been attributed to the lack of alkali normally secreted in the parotidean saliva, which is needed to neutralize acids formed in the animal's stomach.

The principal functions of saliva are (*a*) dilution and lubrication of foods, so that chewing and swallowing are facilitated; (*b*) the solution of sapid substances, so they can stimulate the taste buds; (*c*) moistening and cleaning of the mouth and protection of the teeth; (*d*) the commencement of the digestion of starch.

Claude Bernard[1] attributed to the saliva of each gland a special function: submaxillary saliva is stimulated principally by sapid substances and is the saliva of taste; parotidean saliva serves to moisten the mouth and foodstuffs and thus facilitates mastication; sublingual saliva lubricates the food and favors swallowing. This opinion has been confirmed to a great extent.

Imbibition and lubrication of food. One of the most important functions of saliva is to facilitate chewing and swallowing. Saliva moistens dry foodstuffs and thus furthers mastication and the formation of the bolus, covering it with a lubricant layer which helps it to be swallowed.

Action on taste. A sapid substance must be dissolved to act on the taste buds; *e.g.*, quinine, a very bitter but only slightly soluble substance, when placed on the tongue in the form of powder has scarcely any taste. Saliva dissolves solid foods, which can then be tasted.

Lubricant action on the mucosa. Saliva lubricates the mucosa of the mouth and thus facilitates mastication and speech. Subjects suffering from aptyalism must frequently moisten their mouths with water. Public speakers sometimes have to sip water frequently to complement an insufficient salivary flow. Saliva also

[1] BERNARD, C., "Leçons de physiologie expérimentale," vol. 2, J. C. Baillière et fils, Paris, 1856.

protects the teeth; dental caries is common in cases of aptyalism.

Digestive function. Human saliva and that of some herbivora contain an amylase called ptyalin which accelerates the breakdown of starch and glycogen into maltose. Food remains but a short time in the mouth, so ptyalin could have but little effect on carbohydrate digestion, except for the fact that the saliva impregnates the whole bolus, and as the gastric juice takes some time to penetrate into its deeper layers, ptyalin can exercise its hydrolytic effect in the stomach. As the acidity increases, the action of ptyalin diminishes until it ceases completely.

Role of saliva in thirst. When there is a great loss of water (sweating, diarrhea, polyuria, hemorrhage, etc.), salivary secretion diminishes and can cease completely. The subsequent drying of the mouth is one of the principal factors of the sensation of thirst.

MASTICATION

The crushing and breaking up of the food begin in the mouth, by the combined action of the jaws and the teeth, the masticatory muscles, the tongue, and the cheeks.

The articulation of the lower jaw with the skull varies in different species according to their diet. In carnivora the condyle is firmly fixed in the cavity of the temporal bone; a strong joint with limited movements is thus formed. In ruminants there is a small condyle and a shallow glenoid cavity, which permits ample lateral movements. The disposition of the joint in rodents is such that it facilitates movements of projection and retraction of the jaw.

In man this joint is so formed that all kinds of movements can be executed. The muscles that move the jaw can be divided into those that raise it (elevators) and those that lower it (depressors). Each one of these muscles also projects or retracts the jaw, and by the combined action of several muscles lateral movements are performed. The following diagram shows the masticatory muscles and their functions:

Elevators..	Projectors...	Masseter Internal pterygoid
	Retractor...	Temporal
Depressors..	Projector....	External pterygoid
	Retractors...	Digastric Mylohyoid Geniohyoid

The combined action of the muscles of the tongue and cheeks is necessary for efficient mastication; these muscles act as the hopper in the mill (Cannon), placing the food between the dental arches. They also collect the particles of food into a mass or bolus, which is then swallowed.

Just as the temporomandibular joint varies in different species according to their diet, the teeth also vary; rodents have well-developed cutting incisors, carnivora have large canines which are used in tearing the food, and herbivora have large grinding molars. Sherrington[1] and Bremer[2] have studied a series of masticatory reflexes, which continue to take place in orderly sequence even after the cerebral cortex has been removed. Bremer has described three types of reflexes which are dependent on the place of excitation:

1. When the stimulus acts on the incisors, the anterior part of the masseter contracts and produces a series of rapid movements of the jaw. This is the gnawing reflex.
2. When the stimulus acts on the mucosa of the mouth anterior to the molars, ample rhythmic movements which raise and lower the jaw are produced. This is the vertical masticatory reflex.
3. When the stimulus acts on the mucosa near the molars, lateral movements of the jaw are provoked. This is the ruminatory reflex.

Sherrington has shown that stimulation of the mucosa of the mouth by the bolus reflexly depresses the jaw by inhibiting the muscles that raise it. This is followed by a rebound (see Chap. 70), in which these same muscles contract and sharply close the mouth. These movements are repeated rhythmically.

The centers of the masticatory reflexes are situated in the pons and the medulla.

Mastication crushes the food and facilitates its mixture with saliva and other digestive juices. Mastication also helps to stimulate the secretion of saliva.

Insufficient mastication does not seem to have any appreciable effect on the digestion and absorption of food. Nevertheless adequate chewing favors the complete digestion of foodstuffs that contain cellulose.

[1] SHERRINGTON, C. S., *J. Physiol.*, **51**, 420, 1917.
[2] BREMER, F., *Arch. internat. de physiol.*, **21**, 308, 1923.

DEGLUTITION (THE ACT OF SWALLOWING)

The passage of food from the mouth to the stomach is called deglutition.

Methods of study. Several methods have been used to study the act of swallowing in both man and animal. Interesting data have been given by simple observation, especially in cases with surgical or accidental injuries; by the introduction of balloons for registering the variations in pressure; and by auscultation and registration of the sounds produced in swallowing. Observation by means of x-rays is undoubtedly the best method for the study of deglutition.

FIG. 163. X-ray diagrams of the stages in swallowing a liquid. At the beginning of ingestion the tip of the tongue is depressed and the liquid accumulates between the tongue and the dental arches (*a*). Next the tip of the tongue rises and the back part of the tongue is depressed (*b*); the liquid thus runs backward on an inclined plane (*c*). Finally the posterior part of the tongue is sharply drawn upward and backward, sending the liquid into the pharynx (*d*) and (*e*). (*Mosher, H. P., Laryngoscope, vol. 37, p. 235, 1927.*)

Description of the act of swallowing. Since Magendie[1] made his fundamental study of deglutition, it is customary to consider three stages in this act: (*a*) the first stage, which takes place in the mouth; (*b*) the second, in the pharynx; (*c*) the third, in the esophagus.

The first stage is under voluntary control. The soft mass of food, already chewed and mixed with saliva, or the liquids that have been taken into the mouth are placed between the tip of the tongue and the incisors. From there they pass to a hollow formed by the dorsum of the tongue covered by the palate. The tip of the tongue is then raised and pressed against the palate and the upper dental arch; the base of the tongue is depressed. Next the whole anterior part of the tongue is raised against the palate so that the bolus slides backward on an inclined plane. Finally the back of the tongue is sharply moved upward and backward, pushing the bolus into the pharynx (Fig. 163). The contraction of the mylohyoid muscles is the most important movement in this stage; the styloglossi and the glossopalatini assist in drawing the tongue backward.

The second stage in deglutition commences with the arrival of the bolus to the space between the base of the tongue and the posterior wall of the pharynx (Fig. 164). This stage and the following one are no longer under voluntary control, but are purely reflex. The stimuli act on the nerve endings in the mucosa of the pharynx, the epiglottis, and the soft palate; excitation is transmitted by the glossopharyngeal nerve to the center of deglutition in the medulla, where motor response is coordinated, and thus the complex act of swallowing continues.

The second stage is important because, as it passes through the pharynx, food crosses the

[1] MAGENDIE, F., "Précis élémentaire de physiologie," Paris, 1880.

respiratory path. Once in the pharynx the food mass as well as entering the esophagus might travel along one of three other ways: (*a*) it could go back into the mouth; (*b*) it could go up into the nasopharynx; (*c*) it could go down into the larynx. There are adequate mechanisms to stop the food from going in these directions.

Sustained contraction of the muscles responsible for the projection of the bolus into the pharynx prevents its return into the mouth. The contraction of the glossopalatini, which form the faucial pillars, narrows the isthmus of the fauces; the path back to the mouth is completely closed by the base of the tongue. Moreover contraction of the glossopalatini and the pharyngopalatini muscles produces a negative pressure in the pharynx and the esophagus,[1] while there is a positive pressure in the mouth; thus the progress of the bolus is further assisted.

The opening into the nasopharynx is closed by the contraction of the levator and tensor veli palatini and the muscles of the uvula, which raise the soft palate. These muscles also dilate the eustachian tubes and thus regulate the pressure in the middle ear. In cases of paralysis of these muscles (diphtheria, etc.) the act of swallowing is accompanied by partial regurgitation of liquid food into the nasal fossae.

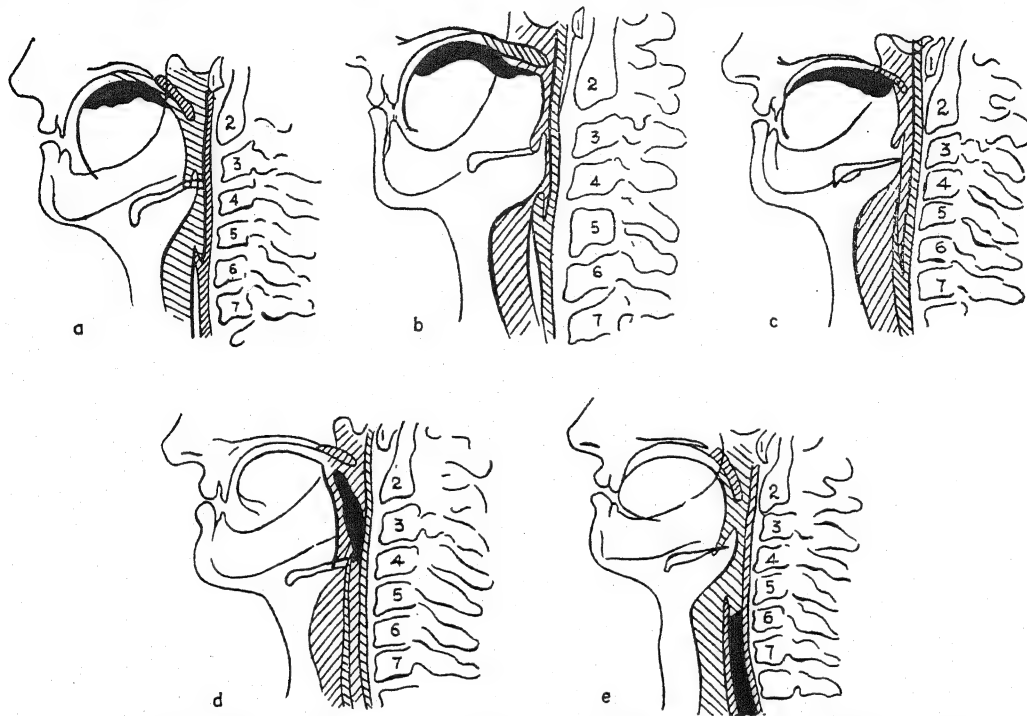



FIG. 164. Stages of swallowing. *a*, when swallowing begins the bolus is pressed between the tongue and the palate; the larynx rises and the pharynx contracts. *b*, the bolus is tightly pressed against the hard palate; the soft palate rises and closes the nasopharynx; the larynx and the pharynx are closed. *c*, the pharynx begins to open, the lower part contracts and keeps the entrance to the larynx closed. *d*, the bolus enters the pharynx and slides over the epiglottis. *e*, larynx and pharynx return to their normal position as the bolus passes down the esophagus. (*Barclay, A. E., Brit. J. Radiol., vol. 3, p. 534, 1930.*)

The upper opening of the larynx is brought under the base of the tongue and closed by the epiglottis by the simultaneous contraction of the thyrohyoid and the suprahyoid muscles.

At the same time the vocal cords are brought together by contraction of the muscles attached to the arytenoid cartilages, which are drawn forward and rotated medially; thus the glottis is closed, and simultaneously the upper end of the esophagus is opened. The important part of this act is the raising of the pharynx. The epiglottis plays a minor part; its upper half can be destroyed without disturbing the act of swallow-

[1] BARCLAY, A. E., *Brit. J. Radiol.*, 3, 534, 1930.

ing. On the other hand, any lesion that hinders the movements of the larynx provokes serious difficulty in swallowing.

All the movements described take place very rapidly. They have been studied by roentgen-cinematography.[1] If the food swallowed is liquid or semiliquid, contraction of the mylohyoids, together with the contraction of other muscles in the floor of the mouth, projects it in a stream through the pharynx into the esophagus. If the food is solid or a soft paste, it also passes quickly through the pharynx, but in addition it is helped by the contraction of the constrictor muscles of the pharynx. The pharynx, which is initially contracted, relaxes suddenly and a negative pressure results; then the median and lower constrictors contract successively, impelling the food downward. At the beginning of the first stage a short inspiration (inspiration of swallowing) takes place, then there is complete apnea until the end of the second stage.

The third stage is the one during which the food progresses along the esophagus to the cardia. The second stage is completed in less than 1 sec.; the third stage is not so rapid. Fluids arrive at the cardia in 1 to 2 sec., and a soft well-lubricated bolus takes a few seconds.

In man progression of a solid or a semisolid bolus is assured by contractions of the esophagus in the form of peristaltic waves. Liquids pass rapidly down to the cardia while the esophagus is totally relaxed; for this reason corrosive fluids do not damage equally the whole length of the esophagus, but mainly its upper end and the part near the cardia.

Gravity helps the descent of fluids and semiliquid foods; but in man swallowing can be carried out even against the force of gravity. It is possible that the bolus is preceded by a wave of negative pressure, which helps to draw the food toward the cardia.

The nervous mechanism of deglutition. Reflex swallowing, *i.e.*, the second and third stages, commences with the application of stimuli on the nerve endings of the mucosa of the anterior and posterior pillars of the fauces, the uvula, the anterior aspect of the soft palate, the posterolateral wall of the lower pharynx, and the epiglottis. Local application of cocaine on these areas makes swallowing impossible for a time. The afferent paths of the reflex are the glossopharyngeal nerve, the upper laryngeal branch of the vagus, and the second division of the fifth cranial nerve.

The center of deglutition is situated on the floor of the fourth ventricle, a little above the respiratory center, in the neighborhood of the nucleus of the vagus. From this center impulses are sent to the motor centers that take part successively in the complex act of swallowing.

The efferent paths of this reflex go along the hypoglossal nerve to the muscles of the tongue—the trigeminal to the mylohyoid muscle; and the glossopharyngeal, the vagus, and the spinal accessory to the muscles of the pharynx, the larynx, and the esophagus. Once the reflex is started the whole series of movements takes place, including the peristaltic waves along the esophagus.

Mosso's[1] classic demonstration showed that the peristaltic waves depended on the integrity of the extrinsic innervation (the vagus nerve). He cut the esophagus into several segments and started a swallowing reflex by stimulating the pharynx. A peristaltic wave was thus started which traveled along the esophagus although there was no anatomical continuity of the gullet.

Section of the vagus provokes paralysis and dilatation of the esophagus and spasm of the cardiac sphincter. Food is not passed along to the stomach, but may go into the respiratory tubes and thus cause death. After some time the cardia relaxes and food can pass into the stomach.

The coordination of breathing and swallowing is of great importance. If breathing did not cease on swallowing, food might pass into the respiratory tubes, and this would have serious consequences. The apnea which commences at the beginning of the second stage of swallowing apparently is due to impulses carried by the glossopharyngeal nerve, because electrical stimulation of the central end of this nerve causes breathing to stop immediately.

THE CARDIA

Around the lower end of the esophagus there are muscular fibers which act as a sphincter. Usually these fibers are in a state of tonic contraction, but they relax on the arrival of food. Sometimes food is detained at the level of the sphincter until a peristaltic wave traveling along

[1] Ardran, G. M., and F. H. Kemp, *Proc. Roy. Soc. Med.*, **44**, 1038, 1951; Rushmer, R. F., and J. A. Hendron, *J. Applied Physiol.*, **3**, 622, 1951.

[1] Mosso, *Moleschotts Mitt.*, **11** (4), 23, 1874.

the esophagus passes it through the cardia. As long as the stomach contents are slightly alkaline, there is frequent regurgitation from the stomach through the cardia into the lower part of the esophagus. As the acidity of the stomach contents increases, regurgitation diminishes. This is due, according to Cannon, to a stronger tonic contraction of the cardiac sphincter provoked by gastric acidity.

The cardia is innervated by the vagus and the sympathetic nerves. Apparently both nerves have excitatory and inhibitory fibers. The effect of electrical stimulation of these nerves varies according to the previous state of the cardiac tonus; however the effect of the vagus is principally inhibitory and that of the sympathetic mainly excitatory. Bilateral cervical vagotomy disturbs the inhibitory mechanism and produces spasm of the cardia due to sympathetic predominance. In some cases of cardiospasm (achalasia) prostigmine and ergotamine can produce a beneficial effect.

BIBLIOGRAPHY

BABKIN, B. P., "Secretory Mechanism of the Digestive Glands," 2d ed., Hoeber, New York, 1950.

BARCLAY, A. E., "The Digestive Tract; A Radiological Study of Its Anatomy, Physiology and Pathology," Cambridge, New York, 1933.

CATEL, W., "Normale und pathologische Physiologie der Bewegungsvorgänge im gesamten Verdauungskanal," Vol. 1, Thieme, Leipzig, 1936.

EMMELIN, N., Paralytic Secretion of Saliva; An Example of Supersensitivity after Denervation, *Physiol. Rev.*, **32**, 21, 1952.

HEIDENHAIN, R., Beiträge zur Lehre von der Speichelabsonderung, *Stud. Physiol. Inst. Breslau*, **4**, 1, 1868.

LANGLEY, J. N., The Salivary Glands, in Schäfer's "Textbook of Physiology," vol. 1, Y. J. Pentland, Edinburgh and London, 1898, p. 475.

PALUGYAY, J., Schlucken, in Bethe's "Handbuch der normalen und pathologischen Physiologie," **3**, 348, Springer, Berlin, 1927.

PAVLOV, I. P., "The Work of the Digestive Glands," trans. by W. H. Thompson, 2d ed., Griffin, London, 1910.

CHAPTER 37

Digestion in the Stomach

THE STOMACH IS NOT an indispensable organ; it has been completely extirpated in man and animals, and although this operation causes certain disturbances in digestion, they are not incompatible with a normal healthy life. In man the most important function of the stomach is the transformation of food into a homogeneous semifluid mass, the chyme. The stomach also acts as a reservoir of food during the process of digestion, the chyme being passed into the duodenum in small quantities. Gastric juice, owing to its high acidity, has bactericidal activity.

FIG. 165. Diagram of method for registering gastric contractions.

GASTRIC MOTILITY

Methods of study. Many methods have been devised for the study of gastric motility: simple inspection, photography or cinematography of the movements of the stomach after laparotomy or through a window in the abdominal wall, the registration of variations in the gastric pressure by means of a balloon placed in the stomach (Fig. 165), the registration of the electrical phenomena of the stomach wall, etc. The study of isolated, living strips of stomach muscle has also been fruitful. The use of x-rays after ingestion of a meal with a substance opaque to x-rays, such as barium, is undoubtedly the method that has given the greatest amount of information on the movements of the stomach in experimental animals and man.

The structure, shape, and position of the stomach. The stomach is a hollow viscus the walls of which are formed by three muscular layers, covered by the peritoneum on the outer surface and by the gastric mucosa, with the muscularis mucosae, on the inner surface. The external muscular layer is formed by longitudinal fibers, the middle layer by circular fibers, and the internal layer of oblique fibers,

which originate around the cardia, sweep down the smaller curvature, and end by joining up with the circular fibers. At the level of the pylorus the circular fibers are thicker and form the pyloric sphincter. The external and middle layers are more important than the internal one.

In man the stomach has the shape of a J, of a reversed L, or of a steer's horn (Fig. 166). The lowest point of the greater curvature is below a line joining the iliac crests in the majority of individuals in the standing position; in some cases the stomach goes well down into the pelvis. These cases are due not to ptosis or dropping of the stomach but to its lengthening, since the fundus remains in contact with the diaphragm.

FIG. 166. Different types of stomach filled with an opaque meal: *a*, stomach of the reversed-L type; *b*, steer-horn type of stomach; *c*, J type of stomach.

For descriptive purposes the stomach can be divided into two parts: the body and the pylorus, separated by the incisura angularis (Fig. 167). The part of the stomach above the cardia is called the fundus and is usually full of air. The pyloric part comprises the antrum or pyloric vestibule and the pyloric canal. This canal is surrounded by the pyloric sphincter and ends in the duodenal cap or bulb, which is the first part of the duodenum.

The tonus of the stomach. In the empty stomach the cavity is obliterated by the apposition of its walls except at the level of the fundus, which is distended by gas. The arrival of food in the stomach separates its walls, and as the viscus is distended the muscular layers relax, adapting the volume of the cavity to its contents, without any increase, or only a transient one, in the intragastric pressure. This postural tone, common to all hollow viscera, is maintained by a special adaptative mechanism (receptive relaxation). Swallowing causes relaxation of the gastric musculature.

Movements of the stomach. The activity of the fasting stomach passes through three main phases (Fig. 168): (*a*) active contractions, which last 30 to 40 sec. and take place at intervals varying from 10 sec. to 5 min.; (*b*) rhythmic changes in tonus; (*c*) relative quiescence. Usually after a period of activity, which lasts 30 to 60 min., there is another of quiescence, followed by a stage of rhythmic tonus.[1] The

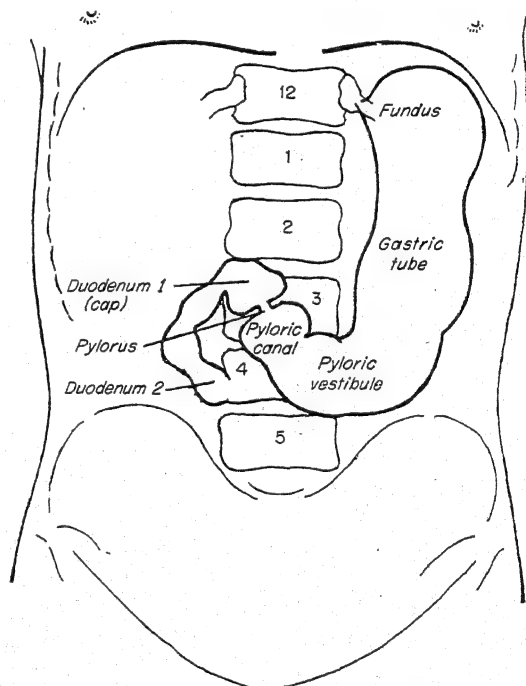



FIG. 167. Diagram of the subdivisions of the human stomach.

[1] ANDERSON, W. F., *Lancet*, 1, 40, 1943.

active periods are usually accompanied by a sensation of hunger,[1] which may be painful.[2]

Almost immediately after the ingestion of food, a peristaltic wave commences in the middle of the body of the stomach, increases in strength at the level of the incisura, and ends in the neighborhood of the pylorus. The upper part of the body of the stomach and the fundus seldom show peristaltic activity.

FIG. 168. Principal types of activity of the fasting stomach: *a*, active contractions; *b*, variations in tonus; *c*, relative quiescence. (*Anderson, W. F., Lancet, vol. 1, p. 40, 1943.*)

Peristaltic waves follow each other rhythmically every 15 to 20 sec.; they take from 15 to 30 sec. to travel down the stomach, and they can be of several types. Sometimes on arriving at the antrum they provoke a systolic contraction of the circular fibers of the whole antrum. At other times peristaltic waves go all the way to the pyloric sphincter. Some of the waves die out soon after they have started. When the pylorus is closed the food flows back into the stomach; the repetition of this maneuver kneads the food and mixes it with the gastric juice.

These movements are kept up as long as there is food in the stomach. The body of the stomach functions as a reservoir; food can remain there for a long time without change. The antrum has a double function: it mixes the contents (the pyloric mill) and evacuates the stomach. The different foodstuffs are mixed with each other and with the digestive juices, being transformed into a semiliquid paste called the chyme. Periodically the pylorus opens and a portion of chyme is passed into the duodenum. Peristaltic waves are not indispensable for the evacuation of the stomach; if the intragastric pressure is greater than the duodenal pressure and the pylorus is open, the chyme flows into the duodenum.

Innervation of the stomach. Peristaltic movements of the stomach can be observed in the denervated stomach and in stomachs that have been extirpated and kept alive by perfusion. Therefore this type of contraction is not dependent on the impulses carried by the vagus and the sympathetic nerves. The walls of the stomach, like those of the rest of the alimentary canal, contain two nerve plexuses, those of Auerbach and Meissner, which undoubtedly play a part in the coordination of the tonic and rhythmic contractions of the different parts of the viscus. According to Cannon, the peristaltic contractions of the antrum are myogenous, as they persist after Auerbach's plexus has been destroyed.

Stimulation of the vagus or of the sympathetic produces variable results on the stomach and on the pyloric sphincter. The effect of the vagus on the musculature of the wall and the pyloric sphincter is predominantly excitatory, and that of the sympathetic is predominantly inhibitory. Nevertheless the results of stimulation depend on the condition of the stomach muscle at the

[1] CANNON, W. B., and A. WASHBURN, *Am. J. Physiol.*, **29**, 441, 1912.

[2] CARLSON, A. G., "The Control of Hunger in Health and Disease," University of Chicago Press, Chicago, 1916.

moment. Hence, if there is a high tonus or the muscle is actively contracting, stimulation of either nerve produces relaxation. On the other hand if the muscle is relaxed and quiescent, stimulation of either nerve increases tone and awakens peristalsis. The immediate effect of double vagotomy is the loss of tone and motility, but soon both are recovered. Sympathectomy does not produce any notable change, although sometimes there is an immediate but transient increase in motility.

Acetylcholine is released on stimulation of the vagus, and sympathin on stimulation of the sympathetic. These substances are chemical mediators of stimulation that act directly on the effectors. Acetylcholine and other choline esters increase the tone and motility of the stomach. The same effect is obtained with eserine and prostigmine, drugs that inhibit cholinesterase. Atropine and adrenaline inhibit the movements of the stomach.[1]

The pyloric sphincter. The pyloric sphincter is formed by the distal segment of the circular fibers of the pyloric part of the stomach. Between the sphincter and the muscular layer of the duodenum there is a barrier of connective tissue, which is bridged by only a few muscle fibers. The pyloric muscle is autonomous and contracts after complete denervation. Nevertheless both the vagus and the sympathetic have a certain action on the sphincter, the former mainly excitatory, the latter mainly inhibitory. The pylorus responds to impulses originated in the hypothalamus; for this reason, in nervous and emotional states signs of pyloric spasm are sometimes observed.

The pylorus (Greek πυλη, door, and οὐρος, keeper) has been considered the sole regulator of gastric evacuation. This opinion can no longer be maintained, as after extirpation of the pylorus or a gastroenterostomy, the emptying of the stomach takes place with no significant alteration. Moreover, during gastric evacuation the pylorus is open most of the time.[2]

When the stomach is empty, the pylorus remains open with an occasional rhythmic contraction. After the ingestion of liquid or semiliquid substances gastric evacuation begins immediately and is soon ended. Solid food remains in the stomach until it is converted by the gastric juice and the kneading action of the antrum into a semiliquid paste. Only then does the stomach begin to empty, a process which takes 3 to 4½ hr. according to the nature of the meal. Carbohydrate is evacuated more rapidly than meat, and fat remains even longer in the stomach (Fig. 169).

The mechanism of gastric evacuation. There is some discussion as to the mechanism of gastric evacuation. At first it was believed that the peristaltic waves forced the stomach contents into the duodenum, if the pylorus was open. There is seldom a contraction sufficiently strong to separate the antrum from the rest of the stomach; the intragastric pressure is therefore not often modified to any considerable degree by peristaltic contractions. More recent work has shown that the most important factor in the evacuation of the stomach is the difference between the intragastric and the intraduodenal pressures; the contents of the stomach are evacuated when the gastric pressure is higher than the duodenal pressure and the pylorus is open. The gastric pressure rises above the duodenal pressure (*a*) when there is increase of gastric tonus or contraction of the whole stomach; (*b*) when there is relaxation of the duodenum; (*c*) when both these conditions occur simultaneously. Any agent that diminishes gastric tone or motility, or increases the duodenal tone, will diminish, annul, or reverse the gradient of pressure. It will therefore retard gastric evacuation or even cause regurgitation of the duodenal contents into the stomach.

Distention of the stomach by the ingestion of food stimulates its motility; intragastric pressure increases, and the evacuation of the stomach is facilitated. The evacuation time of the stomach is not significantly modified by the condition of the pylorus.[1] The tonus and peristaltic movements of the stomach are strongly influenced by duodenal factors. Distention or mechanical irritation of the duodenum, hypertonic or hypotonic solutions, and HCl in the duodenum inhibit gastric motility and lengthen the evacuation of the stomach. Fats[2] and carbohydrates[3] have the same effect; they act reflexly and by

[1] Anderson, W. F., and N. Morris, *J. Pharmacol. & Exper. Therap.*, 77, 258, 1943.

[2] Crider, J. O., and G. E. Thomas, *Am. J. Digest. Dis. & Nutrit.*, 4, 295, 1937–1938.

[1] Crider and Thomas, *loc. cit.*

[2] Quigley, J. P., *Am. J. Digest. Dis. & Nutrit.*, 1, 425, 1934.

[3] Quigley, J. P., and K. R. Phelps, *Am. J. Physiol.*, **109** 133 1934.

means of a humoral mechanism (enterogastrone). Products of protein digestion (proteoses, peptones, and amino acids) act reflexly through the vagus nerves.[1]

Quigley[2] describes gastric evacuation in the following words: "A small portion of the gastric contents enters the intestine. If the sample proves satisfactory, the remainder leaves the stomach rapidly; unsuitable material initiates reactions from the duodenum which temporarily retard further gastric evacuation."

Fig. 169. Time taken to evacuate different foods from the stomach. Except where the amount is specified, the diagram refers to 100-gm. portions. (*Hawk and Bergeim, "Practical Physiological Chemistry," 11th ed., The Blakiston Company, Philadelphia, 1937.*)

The emptying time of the stomach. From what has been said it can be deduced that the principal factors that have an influence on the emptying time of the stomach are (*a*) the intragastric pressure, determined by the tonus and the rhythmic contractions of the stomach muscles; (*b*) the pressure in the duodenal cap, governed by factors similar to those acting on the stomach; (*c*) the state of contraction or relaxation of the pyloric sphincter. The three factors are mainly dependent on stimuli arising in the duodenum.

Water and certain liquid or semiliquid substances pass almost immediately into the duodenum, even when the stomach does not contract; the difference in pressure normally existing between the stomach and the duodenum is sufficient to cause this. However, if the fluid is strongly acid or alkaline, or if it is hypertonic, it takes longer to leave the stomach, because as soon as some enters the duodenal cap it stimulates the closure of the pylorus; for evacuation to continue, the fluid now in the duodenum must be neutralized or diluted.

Table 31. Composition of Food and Gastric Evacuation Time

Composition	*Evacuation Time*
Normal: 2 scrambled eggs with butter, bread, coffee, and milk; total fat, 50 gm.	6 hr.
Fat-poor: 2 boiled eggs, bread, coffee	4 hr., 45 min.
Fat-rich: 2 scrambled eggs with butter, bread, coffee, milk, and 100 gm. butter	After 6 hr. only 60 per cent had been evacuated

Source: Wolf, S., and H. G. Wolff, "Human Gastric Function," Oxford, New York, 1943.

The time of evacuation of solid foods depends on the speed with which they are transformed

[1] Thomas, J. E., and J. O. Crider, *Am. J. Physiol.*, **126**, 28, 1939.

[2] Quigley, J. P., *Arch. Surg.*, **44**, 414, 1942.

into a semiliquid mass, which does not stimulate the duodenum as solids do. Carbohydrates are evacuated more rapidly than proteins; fats remain a longer time in the stomach (Table 31).

Factors which modify gastric tonus and motility condition the evacuation time of the stomach. This time is shorter in hypertonic (steer-horn-shaped) than in hypotonic (J-shaped) stomachs.[1] Rapidity of evacuation seems to be directly related to the volume of the gastric contents; distention of the stomach reflexly increases its tonus and motility.[2]

Enterogastrone (see Chap. 35), secreted by the duodenum when stimulated by fats, fatty acids, sugars, etc., inhibits gastric secretion and motility. Urogastrone has the same effect. Insulin increases gastric motility. Hypoglycemia produced by this hormone stimulates hypothalamic centers, and motor impulses are conducted by the vagus nerves to the stomach. Section of both vagi suppresses the effect of insulin on gastric motility.

Alcohol, caffeine, and choline increase gastric motility, and tobacco, atropine, and banthine[3] (see "Effects of certain drugs on gastric secretion," page 359) depress it.

VOMITING

Vomiting is a complicated act which produces the ejection of the gastric contents through the mouth. Usually vomiting is preceded by a sensation of nausea, which coincides, according to Barclay, with a decrease in the tonus of the stomach; there is also excess salivation and the respiratory rhythm is modified. Retching follows these preliminary phenomena. This is due to uncoordinated movements of the respiratory muscles; the diaphragm and the expiratory muscles contract simultaneously.

Vomiting begins with a deep inspiration, the glottis is closed, and the soft palate is raised so as to close the nasopharynx. The diaphragm and the abdominal muscles contract strongly. The cardia and the esophagus relax; the body of the stomach also relaxes. Frequently there is a strong peristaltic wave, which sometimes divides the stomach into two cavities at the level of the incisura. All these movements are combined so as to expel the stomach contents. The small amount of food that remains in the esophagus after vomiting has ceased is evacuated by the increase in thoracic pressure, which results from the relaxation of the diaphragm together with the contraction of the expiratory muscles, and by antiperistaltic movements in the esophagus.

At one time it was thought antiperistaltic contractions of the stomach were of great importance in the mechanism of vomiting. However, antiperistalsis is not observed; on the contrary, the wave of contraction that starts in the incisura frequently travels down to the pylorus as in the normal movements of the stomach. Magendie demonstrated, over a century ago, that the stomach musculature does not play an essential part in vomiting. He replaced the stomach of a dog by a hog's bladder and then provoked vomiting, which ejected the contents of the bladder through the mouth.

The increase in intragastric pressure, produced by the contractions of the diaphragm and the abdominal muscles, would be ineffectual if the cardia and the esophagus did not relax at the same time. Defecation and coughing are also accompanied by a great increase in intra-abdominal and therefore intragastric pressure, but there is no vomiting because the cardia does not relax. Some animals, such as the horse, the rabbit, and, generally speaking, the herbivora, do not vomit. In some instances, *e.g.*, in the horse, this seems to be due to the lack of relaxation of the cardia; ineffectual vomiting movements occur, which are occasionally so strong that the stomach bursts. In other animals there is no coordinating center for vomiting movements.

The nervous mechanism of vomiting.[1] According to Hatcher, vomiting is always a reflex movement, which can start in many different parts of the body. Afferent pathways carry the impulses to the center in the medulla, whence coordinated impulses are sent out to the muscles.

The following stimuli can provoke vomiting:

1. Mechanical stimulation of the isthmus of the fauces. The afferent path is the trigeminal or the glossopharyngeal.
2. Irritation of the stomach by food or drugs: tartar emetic, mustard, ipecac, mercuric

[1] Macht, T. H., M. H. F. Friedman, and B. H. Malone, *Federation Proc.*, **6**, 161, 1947.

[2] Hunt, J. N., and W. R. Spurrell, *J. Physiol.*, **113**, 157, 1951.

[3] Longino, F. H., *et al.*, *Gastroenterology*, **14**, 301, 1950.

[1] Hatcher, R. A., *Physiol. Rev.*, **4**, 479, 1924.

chloride, zinc sulfate, etc. The afferent path is constituted by fibers included in the vagus, the sympathetic, or both nerves, according to which agent is present.

3. Distention, compression, or irritation of the abdominal viscera: intestine, appendix, bile ducts, uterus, kidney, etc.
4. Dilatation of the heart and other stimuli, such as digitalis and pilocarpine, which act on this organ. The afferent impulses travel along the cardiac nerves.

These afferent impulses are integrated in the "vomiting center," a localized area in the reticular formation of the medulla,[1] the destruction of which makes vomiting impossible. Borison and Wang have also located an area on the ala cinerea of the fourth ventricle which is sensitive to apomorphine and digitalis glucosides.[2] These substances do not act directly on the vomiting center but on chemoreceptors in the fourth ventricle, and perhaps on peripheral chemoreceptors. Efferent impulses of the reflex travel (*a*) along the vagus to the stomach; (*b*) along the phrenic nerves to the diaphragm; (*c*) along the spinal nerves to the abdominal muscles. There are also inhibitory impulses which go from the center to the cardia and esophagus.

The excitability of the center varies in different individuals, and in the same individual in different circumstances. Hatcher maintains that the center is constantly receiving impulses from all parts of the body, which do not result in vomiting because they cause only subliminal excitation. When the excitability of the center increases—*i.e.*, when the threshold is lowered, as in certain psychic states, in fatigue, and by some drugs (apomorphine, emetine, etc.)—the usual afferent impulses are enough to cause vomiting. On the contrary if the excitability of the center is lowered, a previously efficacious stimulus no longer provokes vomiting. This decrease in excitability is obtained by increasing the activity of neighboring centers and by giving certain drugs (barbiturates, etc.). Nausea and vomiting form part of the syndrome of vertigo and seasickness.

[1] Borison, H. L., and S. C. Wang, *J. Neurophysiol.*, **12**, 305, 1949; Wang, S. C., and H. L. Borison, *Arch. Neurol. & Psychiat.*, **63**, 928, 1950.

[2] Borison, H. L., and S. C. Wang, *Proc. Soc. Exper. Biol. & Med.*, **76**, 335, 1951, and **77**, 38, 1951.

GASTRIC SECRETION

The gastric glands. The gastric mucosa, from the cardia to the pylorus, has many glands. These are of three different types. Around the cardia there are composite tubular glands, the cardiac glands. In the pyloric region there are short glands with a long and sinuous excretory duct, the pyloric glands. In the mucosa of the body or fundus of the stomach, the peptic or principal glands are found; they have straight excretory ducts which end in the fovea of the mucosa. The pyloric glands resemble the Brünner glands of the duodenum; they secrete an alkaline juice rich in mucus. There are three types of cells in the principal glands: (*a*) the chief, zymogenic, or peptic cells, which are cylindrical in shape, have poorly marked contours, and are mostly found in the body of the gland; (*b*) the parietal, border, or oxyntic cells, which are of varying shape, have well-marked contours, are placed between the basal membrane and the chief cells, and are more numerous in the neck of the gland; (*c*) mucous cells, which are found in the neck of the gland. The chief cells secrete pepsinogen, the constituent substance of the cytoplasmic granules; these granules decrease in amount when the cells secrete. The parietal or border cells secrete hydrochloric acid and for this reason are also known as oxyntic cells.

Gastric Juice

Gastric juice is a mixed secretion from the tubular glands of the stomach. Gastric juice must be distinguished from the gastric contents, which is the small amount (30 to 50 cc.) of fluid made up of gastric juice, mucus, saliva, and duodenal content that is found in the fasting stomach or that can be removed from the stomach after a meal.

Methods of collecting gastric juice. In man the contents of the stomach can be obtained by a catheter. The contents are removed for analysis 1 to 2 hr. or every 15 min. after a test meal.

Analysis of human gastric juice. A procedure usually followed is to (*a*) introduce a special catheter into the stomach; (*b*) draw out the contents of the stomach; (*c*) give a test meal (one commonly used is that of Ewald-Boas, which consists of 250 cc. of weak tea, 35 gm. of toasted bread, and 10 gm. of sugar); (*d*) withdraw from 5 to 6 cc. of the contents every 15 min. until the stomach is empty.

Total acidity, free acidity, pH, peptic activity, etc., can be determined in the samples. Total and free acidity are measured with a 0.1 *N* solution of Na(OH) and appropriate indicators. Free acidity is determined by the use of indicators that change color at pH 3 to 4.6 (bromophenol blue, etc.); total acidity by the use of indicators that change color at pH 8.3 to 10 (phenolphthalein, etc.). Acidity is measured in terms of the amount of 0.1 *N* Na(OH) necessary to neutralize 100 cc. of gastric juice, or the amount of HCl that corresponds to 100 cc. of gastric juice, *i.e.*, the percentage of hydrochloric acid (1 cc. of 0.1 *N* HCl contains 0.00365 gm. of HCl) (Fig. 170).

FIG. 170. Typical curves of gastric acidity following a test meal. Solid line, total acidity; broken line, free hydrochloric acid. Scale in minutes.

Pure gastric juice can be obtained by first washing out the stomach and then injecting subcutaneously 0.01 mg. of histamine per kilogram of body weight. This procedure normally produces about 200 cc. of gastric juice in 1 hr. The secretion so obtained is not normal gastric juice because histamine stimulates mainly the secretion of hydrochloric acid and to a lesser degree the secretion of mucus and pepsin.

Cases of accidental or surgical fistulas of the stomach have been of great use in the study of gastric secretion in man.

Between the years 1825 and 1833 William Beaumont[1] made a series of observations and experiments on a Canadian hunter, Alexis St. Martin, who had a gastric fistula as a consequence of having been accidentally shot in the abdomen three years before. St. Martin was nineteen years old at the time of the accident and he lived with his fistula to the age of seventy-eight. In 1876 gastrostomy in a case of cancer of the esophagus was performed for the first time; Richet studied this patient and confirmed many of Beaumont's observations. Later many similar studies on patients with gastric fistula were published. Those of Carlson[2] and Wolf and Wolff[3] are particularly notable for the wealth of detail and care with which they have been made.

Pavlov[4] developed the technique of obtaining

[1] BEAUMONT, W., "Experiments and Observations on the Gastric Juice and the Physiology of Digestion," Plattsburg, 1833.

[2] CARLSON, A. J., "The Control of Hunger in Health and Disease," University of Chicago Press, Chicago, 1919.

[3] WOLF, S., and H. G. WOLFF, "Human Gastric Function. An Experimental Study of a Man and his Stomach," Oxford, New York, 1943.

[4] PAVLOV, I. P., Die äussere Arbeit der Verdauungsdrüsen und ihr Mechanismus, in Nagels, "Handbuch der Physiologie des Menschen," 2, 666, Friedrich Vieweg & Sohn, Brunswick, 1907.

gastric juice in the dog. To avoid contamination of the gastric secretion with saliva and the products of digestion, Pavlov combined a gastric fistula with one in the esophagus (Fig. 171) and improved Heidenhain's gastric pouch. Pavlov's operation consists in completely separating a part of the stomach by a layer of the mucosa, leaving a bridge of the muscular layer and the peritoneum so that the vascular and nervous connections remain intact (Fig. 173).

FIG. 171. A dog with fistulas in the esophagus and stomach, held in Pavlov's holder.

The properties of gastric juice. The amount of gastric juice secreted in 24 hr. varies in different circumstances. The fasting stomach of man secretes 8 to 15 cc. of gastric juice per hour. The composition, acidity, and digestive potency vary considerably, because gastric juice is a mixture of the secretions of several types of gland which respond differently to different stimuli. The isotonic solution of HCl secreted by the parietal cells is diluted and in part neutralized by the secretions of other cells in the gastric mucosa. Gastric juice is a clear, colorless fluid, without smell and strongly acid (pH 0.9 to 1.5). The specific gravity is 1.006 to 1.009 and the freezing point −0.55 to −0.62°C. The principal components are hydrochloric acid, mucus, and enzymes; the enzymes are pepsin, cathepsin, rennin, and a very small amount of lipase.

The activity of gastric juice. The *mucus* in gastric juice protects the mucosa from irritating substances, neutralizes the acid in the gastric content, and forms a precipitate when the acid is in high concentration.

The digestive activity of gastric juice is due in great part to *pepsin.* The principal cells of the glands in the fundus secrete *pepsinogen,* which is converted into pepsin by the HCl secreted by the parietal cells of these same glands. Pepsin is a proteinase, *i.e.*, an enzyme that acts on proteins of high molecular weight; it hydrolyzes the peptide bonds and transforms most proteins into proteoses and peptones. It has a maximum activity at pH 1.5 to 2. This high acidity in the stomach is due to the secretion of hydrochloric acid. Nevertheless in the first 2 hr. after a meal the pH of the gastric contents is between 5 and 6, because of the buffer action of the proteins in the food. Pepsin is inactive at this pH. In other words, pure gastric juice has active pepsin but no substrate on which this enzyme can act; on the other hand the gastric content after a meal has abundant substrate, but pepsin is inactive. A second proteolytic enzyme has been discovered in gastric juice, which has the properties of a *cathepsin.* This enzyme acts on proteins of high molecular weight, just as pepsin does, but its optimum pH is between 3 and 5. According to Buchs[1] cathepsin and pepsin are two enzymes. At the beginning of digestion, while the gastric content has a pH of 3 to 5, cathepsin only is active; later, in the third hour after the ingestion of food, the pH falls to a level at which pepsin becomes active. In brief, cathepsin commences proteolysis in the stomach and pepsin ends it.

Hydrochloric acid not only produces an adequate medium for the activity of pepsin but also has an important digestive effect; it dissolves and disintegrates nucleoproteins, dissolves collagen, precipitates the caseinogen of milk, hydrolyzes sucrose into glucose and fructose, etc.

[1] BUCHS, S., "Die Biologie des Magenkathepsins," ed. Karger, Basel, New York, 1947.

An antiseptic function has also been attributed to HCl; the acidity of gastric juice prevents the development of bacteria of putrefaction and lactic fermentation, and of several pathogenic microbes. Hydrochloric acid also plays a part in the formation of secretin.

Rennin (*Labferment*, *présure*, *cuajo*) is also a protease found in gastric juice. It is formed from a zymogen by the action of acid. This enzyme is found in large amounts in the fourth stomach of calves; it activates hydrolysis of the caseinogen (casein) of milk and transforms it into casein (paracasein). Calcium is necessary for this process, which results in the coagulation (curdling) of milk due to the precipitation of calcium caseinate. The optimum pH for rennin activity is 5.35. Tauber and Kleiner's[1] work has shown that rennin and pepsin are two different enzymes, and thus a long-standing discussion as to the possible identity of these enzymes has come to an end. Nevertheless pepsin and other proteolytic enzymes can coagulate milk, and the existence of rennin in human gastric juice and in the stomach of other mammals has not been definitely proved.

The principal function of gastric juice in the digestion of fats is to dissolve the protein membranes of adipose cells. It also contains a weak *lipase* which disintegrates alimentary fats already emulsified (egg yolk, etc.). Its effects are produced mostly at the beginning of gastric digestion, when the acidity is not too high. The optimum pH is around 6, varying somewhat in different species.

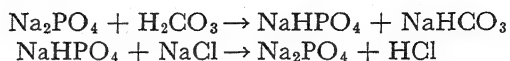
Gastric juice contains an *intrinsic factor*, which acts on an *extrinsic factor* found in several foodstuffs and forms the *hematinic principle* or *antianemic factor*, which is necessary for the normal development of red blood cells. The intrinsic factor is neither hydrochloric acid nor any of the enzymes already mentioned (see Chap. 5).

The formation of hydrochloric acid. The parietal cells secrete hydrochloric acid at a constant concentration. Their product of secretion seems to be an isotonic solution of HCl, which contains 166 mEq./liter Cl^-, $159H^+$, and $7K^+$.[2] The parietal cells can concentrate H^+ to a million times the value of its concentration in the blood. This extraordinary process

[1] TAUBER, H., and I. S. KLEINER, *Ztschr. f. physiol. Chem.*, 220, 205, 1933; KLEINER, I. S., and H. TAUBER, *J. Biol. Chem.*, **106**, 501, 1934.

[2] GRAY, J. S., *Gastroenterology*, **1**, 390, 1943.

has been the object of much study. One of the theories put forward to explain it, which has wide acceptance, is that carbonic acid acts on intracellular disodium phosphate and forms monosodium phosphate, which with sodium chloride gives hydrochloric acid and disodium phosphate.

$$Na_2PO_4 + H_2CO_3 \rightarrow NaHPO_4 + NaHCO_3$$
$$NaHPO_4 + NaCl \rightarrow Na_2PO_4 + HCl$$

Carbonic acid is formed from the CO_2 in the blood by the action of carbonic anhydrase.

The following facts support this theory: (*a*) there are large amounts of carbonic anhydrase in the gastric mucosa; (*b*) an increase in the partial pressure of CO_2 in the blood plasma causes an increase in HCl concentration in the gastric juice;[1] (*c*) the parietal cells contain an exceptionally large quantity of phosphates.

The following alternative hypotheses as to the genesis of HCl have been put forward: (*a*) intracellular hydrolysis of NaCl;[2] (*b*) intracanalicular secretion of KCl, with subsequent reabsorption of K, which is replaced by hydrogen ions produced by cellular metabolism;[3] (*c*) secretion of an organic acid, probably pyruvic acid, into the canaliculi, with reabsorption of the acid radicals and replacement by Cl ions;[4] (*d*) release of H^+ by the oxyntic cells using energy obtained from the metabolization of glucose.[5] Recent work has shown that HCl is not formed when there is lack of thiamine or nicotinic acid, vitamins which play a part in the production and destruction of pyruvic acid.

THE MECHANISM OF GASTRIC SECRETION[6]

The normal resting stomach secretes continuously small amounts of gastric juice (basal secretion of the gastric mucosa).[7] The ingestion of food increases the flow of gastric juice. The ac-

[1] BERNARD, C., "Leçons de physiologie opératoire," Paris, 1879; ADAMS, W. L., C. S. WELCH, and B. B. CLARK, *Am. J. Physiol.*, **139**, 356, 1943; KURTZ, L. D., and B. B. CLARK, *Federation Proc.*, **5**, 188, 1946.

[2] HOLLANDER, F., *Gastroenterology*, **1**, 401, 1943.

[3] CONWAY, E. B., and T. BRADY, *Nature*, **159**, 137, 1947.

[4] BULL, H. B., and J. S. GRAY, *Gastroenterology*, **4**, 175, 1945.

[5] DAVIES, R. E., N. M. LONGMUIR, and E. E. CLARK, *Nature*, **159**, 468, 1947.

[6] IVY, A. C., *Surgery*, **10**, 861, 1941.

[7] LIM, R. K. S., and H. C. HOU, *Proc. Soc. Exper. Biol. & Med.*, **26**, **270**, **1929**.

tivity of the different types of cell which make up the gastric glands is conditioned by nervous, hormonal, and chemical factors, which act to stimulate or inhibit each group of cells. The properties of the juice secreted vary with the stimulus and the response that it provokes. For example, appetite secretion and the secretion produced by sham feeding or a real meal are rich in pepsin. Injection of acetylcholine, pilocarpine, and insulin also provokes the secretion of a gastric juice rich in pepsin.

Two periods must be considered in studying gastric secretion: the resting period and the digestive period.

SECRETION DURING REST

The amount of gastric juice secreted between the digestive periods is usually small, and the concentration of pepsin and HCl is variable; frequently no HCl is secreted. Pavlov believed that a psychic stimulus was necessary for the secretion of HCl in this condition, but more recently it has been found that such a stimulus is not indispensable, at least in man (Babkin). The causes of this continuous secretion of gastric juice in the intervals between the digestive periods are not yet known, but probably there are psychic, nervous, and chemical stimuli (Babkin).

SECRETION DURING DIGESTION

There are three phases in the secretion of gastric juice provoked by a meal: the *cephalic*, the *gastric*, and the *intestinal*, so called because of the site of stimulation.

The cephalic phase. The stimuli responsible for this phase act on receptors situated in the head. The contribution of Pavlov and his pupils to the knowledge of this phase has been so important it could well be called the "Pavlov phase."

In a hungry dog with a gastric fistula (Fig. 171) it is possible to demonstrate that the mere sight of food provokes a flow of gastric juice within a few minutes. The smell of food or any other stimulus that has been naturally or artificially associated with the ingestion of food produces the same effect. This "psychic secretion" or "appetite secretion"[1] has also been demonstrated in man, being similar to the secretion of saliva provoked by these same stimuli. Gastric secretion is due in these cases to a "conditioned" reflex. Gastric secretion can also be produced by suggestion in a hypnotized subject. On the other hand, an intense emotional state inhibits gastric secretion previously provoked by an adequate stimulus.

[1] The term "appetite secretion" should be preferred; the term "psychic secretion" is more appropriate for the flow caused by mental images, hypnosis, etc.

When food is given to a dog with a fistula in the esophagus and another in the stomach (sham feeding),[1] gastric juice commences to flow from the gastric fistula in 10 to 15 min. Food and substances with pleasant taste stimulate the taste buds and reflexly through the vagus provoke gastric secretion. Chewing tasteless substances, such as rubber or sponge, does not have this effect.

Psychic secretion and the secretions of appetite and of sham feeding are of reflex origin. The first two are conditioned reflexes; the cerebral cortex plays an important part in their production. Decerebrate dogs do not secrete gastric juice on being shown food or made to smell it, but they respond to sham feeding.[2] In all these cases the vagus is the efferent path of the reflexes. Section of both vagi or the injection of atropine suppresses psychic or appetite secretion and that following sham feeding.

Electrical stimulation of the vagus provokes secretion of gastric juice as a result of the activity of the different types of glandular cells in the stomach. Uvnäs[3] has shown that intravenous injection of an extract of pyloric mucosa increases the secretory effect of excitation of the vagus, and that this effect diminishes if the pyloric region of the stomach is tied off or removed. Stimulation of the vagus has, therefore, a twofold effect: it stimulates the peptic glands directly, and it stimulates secretion of gastrin by the mucosa of the pyloric region, thus stimulating the peptic glands indirectly through a hormonal mechanism. Hormonal and nervous mechanisms thus act synergically.

The gastric phase. The stimuli that provoke gastric secretion during this phase act on the stomach.

Food is given to a dog with a Pavlov pouch and with both vagi cut so as to eliminate the cephalic phase; 20 to 45 min. later the miniature

[1] The food is chewed, tasted, and swallowed but is lost through the fistula in the esophagus.

[2] Zeljony, G. P., XI Internat. Physiol. Congr., Edinburgh, 1923.

[3] Uvnäs, B., *Acta physiol. Scandinav.*, **4**, Suppl. 13, 1942.

stomach begins to secrete. This is not a reflex secretion, since the portion of the stomach deprived of all its nervous connections[1] or transplanted to the region of the mammary gland[2] (Fig. 172) secretes gastric juice when food is introduced into the part of the stomach remaining *in situ*.

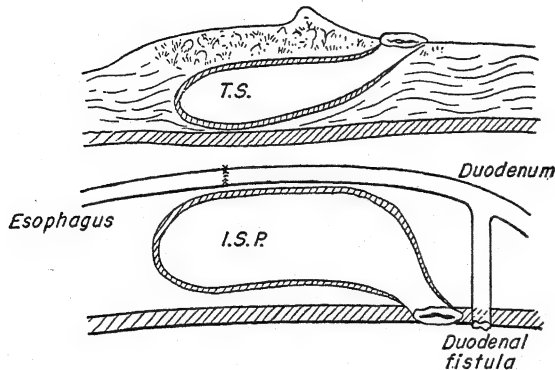



FIG. 172. Ivy's two-pouch dog with a duodenal fistula. *T.S.*, a portion of the stomach disconnected from the principal part and transplanted under the mammary gland; *I.S.P.*, stomach pouch disconnected from the rest of the digestive tract. (*Redrawn from Ivy, A. C., Surgery, vol. 10, p. 861, 1941.*)

The presence of food in the stomach stimulates the secretion of gastric juice. This fact can be demonstrated in a dog with a Pavlov pouch and a fistula in the principal part of the stomach (Fig. 173). When certain foods are introduced through the fistula directly into the stomach, taking care to avoid all psychic stimulation, gastric juice is secreted by the miniature stomach after an interval of 20 to 45 min.

Local mechanical stimulation. Pavlov observed that mechanical stimulation of the gastric mucosa by the introduction of pieces of sponge or sand, or by rubbing the mucosa with a glass rod or a rubber tube, did not provoke gastric secretion. On the other hand distention of the stomach was an efficient stimulus. A continuous flow of 50 cc. of water through the stomach does not provoke the secretion of gastric juice, but a larger volume (150 to 300 cc.) of water does stimulate secretion.

In the experiment described above, the miniature stomach cannot have been stimulated directly by a mechanical action on its mucosa, as it is empty; but since it still has nervous connections with the main part of the stomach, secretion might be due to a "local reflex." Ivy and Farrell transplanted part of the stomach into the mammary region (Fig. 172). This transplant had an outside opening, but was deprived of all nervous connections with the rest of the organism. Distention of the main part of the stomach did not provoke secretion in the transplanted mucosa; therefore the flow of gastric juice produced by distention of the stomach (see preceding paragraph) is due to local stimulation. The transplanted stomach, however, responds with a flow of acid gastric juice to: (*a*) the introduction of certain foods into the stomach *in situ*, and (*b*) distention of the pyloric region of the stomach *in situ*.[1] These experiments give definite proof that a hormonal factor, which stimulates the gastric mucosa, is released from the stomach during the gastric phase of secretion (see Chemical Stimulation).

FIG. 173. Pavlov's miniature stomach.

Chemical stimulation. Not all foods are capable of provoking gastric secretion when introduced into the stomach. Raw egg white, sugar, starch, and bread do not act as stimuli. On the other hand an abundant flow is produced by raw meat, meat extract, meat juice and broth, dilute alcohol, sodium bicarbonate, lactic acid, etc. The products of gastric digestion also act as stimuli; thus raw egg white, which is inactive, stimulates gastric secretion if it is previously digested with HCl and pepsin.

The pyloric portion of the stomach plays an important part in this phase of gastric secretion; the chemical stimuli of gastric secretion must act on this part of the stomach to produce their

[1] LIM, R. K. S., C. T. LOO, and A. C. LIU, *Chinese J. Physiol.*, **1**, 51, 1929.

[2] IVY, A. C., and J. I. FARRELL, *Am. J. Physiol.*, **74**, 639, 1925.

[1] GROSSMAN, M. I., C. R. ROBERTSON and A. C. IVY, *Am. J. Physiol.*, **153**, 1, 1948.

effect. The experiments with transplanted stomachs referred to above exclude all possibility of a reflex mechanism; therefore a humoral agent must play a part.

The nature of this humoral agent is still under discussion. Several possibilities must be taken into account:

1. The stimulating substances could be absorbed and act directly through the blood stream. There are no proofs of this contention, and the positive results obtained by injection of meat and liver extracts might be due to the histamine they contain.
2. Histamine is considered by some to be the gastric hormone, liberated in the pyloric part of the stomach. If this were so it would act on the fundus after having been reabsorbed.[1]
3. The most probable mechanism by which many chemical substances stimulate gastric secretion when they come into contact with the pyloric mucosa is by provoking the formation of a special hormone, called "gastrin," which passes into the blood stream and provokes the secretion of the glands of the fundus. Edkins,[2] who first formulated this hypothesis, showed that intravenous injection of extracts of the pyloric mucosa have a stimulating effect on gastric secretion. Komarov[3] has prepared extracts of the pyloric mucosa free from histamine and choline. Extracts of the mucosa of the fundus and liver extracts do not produce gastric secretion. Pepsin and trypsin destroy the activity of gastrin; therefore it is of proteid nature.[4] The mechanism by which it is liberated during gastric digestion, absorbed, and conveyed by the blood stream to the fundic glands has not yet been demonstrated.

The intestinal phase. The stimuli that provoke gastric secretion during this phase act on the intestine. Chemical stimuli alone are active; distention of the gut and mechanical irritation of the intestinal mucosa do not produce gastric secretion.

Ivy and his collaborators have studied this phase in dogs with fistulas in the duodenum and

[1] Sacks, J., A. C. Ivy, J. P. Burgess, and J. E. Vandolah, *Am. J. Physiol.*, **10**, 331, 1932.

[2] Edkins, J. S., *J. Physiol.*, **34**, 183, 1906; Edkins, J. S., and M. Tweedy, *J. Physiol.*, **38**, 263, 1908.

[3] Komarov, S. A., *Proc. Soc. Exper. Biol. & Med.*, **38**, 514, 1938; *Rev. canad. de biol.*, **1**, 191 and 377, 1942.

[4] Uvnäs, B., *Acta med. Scandinav.*, **6**, 117, 1943.

the stomach, or with the esophagus joined to the duodenum so that the stomach remains excluded, or with a duodenal fistula and a gastric transplant in the mammary gland (Fig. 172). In all these preparations it is possible to demonstrate that certain foods introduced into the duodenum or the jejunum stimulate gastric secretion. Water, meat, meat extract, meat juice or broth, peptones, soaps, and fatty acids are all active. Foods previously digested by pepsin and HCl are more active than undigested foods. The presence of bile seems to be an important factor in the activity of these substances.[1]

These observations show that the intestinal phase of gastric secretion is dependent exclusively on a humoral mechanism, either because a hormone (an intestinal gastrin) is formed in the duodenum and the jejunum and passes into the blood stream, or because the absorbed products of digestion act as stimuli on the fundic glands.

Inhibition of Gastric Secretion

Gastric secretion is inhibited in several conditions. Thus anoxia diminishes gastric secretion in man,[2] and adrenalectomy produces disturbances in the gastric secretion of the rat, of the type known as achylia.[3] The mechanism underlying these changes in the secretion of the stomach is not well known, but in other cases the cause of inhibition is better understood.

Nervous inhibition. Certain psychic states inhibit gastric secretion, as was mentioned before. Inhibitory fibers have been found in the vagi and the splanchnics. Distention of the gut inhibits gastric secretion, probably by a reflex mechanism.

Humoral inhibition.[4] By means of the same methods used to demonstrate the activity of certain substances that act on the stomach or the intestine and by a humoral mechanism stimulate gastric secretion, it is possible to show that other substances acting on the intestine have an inhibitory effect. Fats (olive oil, cream), concentrated solutions of sugar, or sodium bicarbonate inhibit gastric secretion, diminishing the flow and the pepsin and hydrochloric acid

[1] Beamer, W. D., M. H. F. Friedman, J. E. Thomas, and M. E. Rehfuss, *Am. J. Physiol.*, **141**, 613, 1944.

[2] Hartiala, K., and M. Karvoren, *Acta physiol. Scandinav.*, **11**, 85, 1946.

[3] Tuerkischer, E., and E. Wertheimer, *J. Endocrinol.*, **4**, 143, 1945.

[4] Ivy, A. C., *Bull. New York Acad. Med.*, **20**, 5, 1944.

content. Extracts of duodenal mucosa have been prepared which, when injected into an animal, inhibit gastric secretion.[1] The substance responsible for this effect has been called "enterogastrone." "Hormone" (Greek ὁρμαειν, to excite or awake) is the name given to those chemical messengers which, by means of the blood stream, increase the activity of an organ or a system; and "chalone" (Greek χαλαειν, to slacken) is that given to chemical messengers that have an inhibitory effect. Thus enterogastrone is a chalone, as it inhibits not only the secretion, but also the movements of the stomach. Extracts have been obtained that act only on the secretion but have no effect on the movements of the stomach, so it seems there are two different agents. In the urine of dogs and of man, a substance called "urogastrone"[2] has been found which has the same activity as enterogastrone and which probably is the latter after passing through the kidney. Recently Ivy and his collaborators have obtained a purified and concentrated enterogastrone which, when injected intramuscularly or intravenously, prevents the formation of experimental peptic ulcers in dogs and improves the patient's condition in cases of peptic ulcer in man.

Effects of Certain Drugs on Gastric Secretion

The action of certain drugs on gastric secretion is worthy of mention, as they are of use in the practice of medicine.

Subcutaneous injection of 1 mg. of atropine diminishes or suppresses in the dog the continuous secretion that takes place between the digestive periods and also the secretory response to a meal in all its phases—cephalic, gastric, and intestinal. The same dose diminishes, but does not suppress, the response to histamine. In a normal man injection of 1 to 2 mg. of atropine suppresses the continuous secretion and in some instances the secretory response to a meal. In some apparently normal subjects and in cases of duodenal ulcer, atropine has no effect; in these circumstances gastric secretion might be caused by the release of histamine, which is not inhibited by atropine.

[1] Lim, R. H. S., S. M. Ling, and A. C. Liu, *Chinese J. Physiol.*, **8**, 219, 1934; J. S. Gray, W. B. Bradley, and A. C. Ivy, *Am. J. Physiol.*, **118**, 463, 1937.

[2] Culmer, C. U., A. J. Atkinson, and A. C. Ivy, *Endocrinology*, **24**, 631, 1939; Culmer, C. U., J. S. Gray, J. L. Adkison, and A. C. Ivy, *Science*, **91**, 147, 1940.

Banthine (β-diethyl-aminoethylxanthene-9-carboxylate metabromide) is a salt of a quaternary ammonium which inhibits the effect of acetylcholine released by postganglionic parasympathetic fibers and blocks synaptic transmission in sympathetic ganglia. It has been used successfully in the treatment of gastrointestinal disturbances accompanied by hypersecretion and hypermotility.

Subcutaneous injection of 0.5 to 1 mg. of histamine increases considerably the flow of gastric juice with a high concentration of HCl and poor in pepsin and mucus. This effect is used as a test of gastric secretion in man, determining (*a*) the maximum acidity, *i.e.*, maximum concentration of acid, which in normal subjects is the equivalent of 102 cc. 0.1 *N* HCl solution per 100 cc. of gastric juice; (*b*) the volume of acid secreted per hour, which is normally 182 cc. 0.1 *N* HCl; (*c*) the flow of gastric juice per hour, which is normally 200 cc.

Caffeine, injected intramuscularly or given by mouth, stimulates gastric secretion in man and in the cat, but not in the dog.[1] This effect is not suppressed by vagotomy or the injection of atropine. Caffeine apparently acts not by liberating histamine but directly on the parietal cells.[2]

Sodium bicarbonate has a complex action. It neutralizes in part the acid contents of the stomach. If it is rapidly evacuated into the duodenum, as occurs when it is taken on an empty stomach, it inhibits gastric secretion by its effect on the duodenal mucosa. If it remains in the stomach, on the other hand, it acts on the pyloric mucosa and stimulates gastric secretion; finally it also has a stimulating effect once it has been absorbed from the intestine. Its use in therapy is justified only when an immediate neutralization of the gastric contents is desired.

Silica and aluminum gels are used in the treatment of peptic ulcers. They do not completely neutralize the gastric contents, but they do not stimulate gastric secretion, they are not absorbed, and they inactivate pepsin.

Insulin stimulates gastric secretion when it produces marked hypoglycemia; vagotomy suppresses this effect. The activity of pilocarpine is still under discussion. It is generally admitted that this drug stimulates the secretion of the gas-

[1] Roth, J. A., and A. C. Ivy, *Am. J. Physiol.*, **141**, 454, 1944.

[2] Roth, J. A., and A. C. Ivy, *Gastroenterology*, **5**, 129, 1945.

tric enzymes and mucus, but some workers also report the stimulation of hydrochloric acid secretion.

The Normal Course of Gastric Secretion

So far we have considered analytically the effects of the several stimuli that act on the flow of gastric juice and the experimental methods that serve to demonstrate the different phases of secretion, according to the place where these stimuli act. Now we can examine synthetically the normal course of gastric secretion.

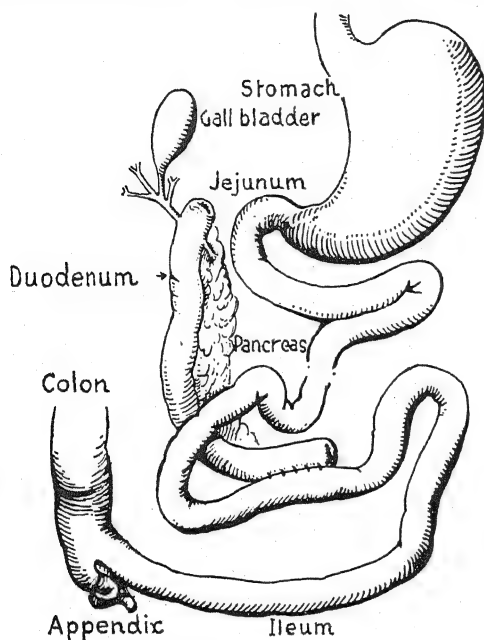



Fig. 174. Diagram of Mann and Williamson's operation for producing experimental ulcers. (*Markowitz, "Textbook of Experimental Surgery," William Wood & Company, Baltimore, 1937.*)

When appetite is aroused (see above, The Cephalic Phase of Gastric Secretion) by the thought of a meal, or by the smell or sight of food, gastric secretion is stimulated by the conditioned-reflex mechanism. On taking food, within a few minutes an abundant flow is produced by way of the vagus nerve; the stimulus in this case is the taste of food. The amount of gastric juice secreted is directly proportional to the pleasure derived from the food. This is the most potent stimulus of gastric secretion, and its importance should be taken into account when considering alimentary hygiene. Food taken with relish because it provokes a pleasurable gustatory sensation will produce an abundant flow of gastric juice and its digestion will thus be furthered. Pavlov very aptly remarked that "appetite is the first and most important stimulus of the secretory nerves of the stomach; in the course of sham feeding, appetite causes the flow of great quantities of highly active gastric juice into the empty stomachs of our dogs. A good appetite at meal times is the source of a copious and active secretion; the absence of appetite means the absence also of gastric juice; the return of appetite assures an abundant secretion from the very beginning of the meal."

Psychic secretion and the reflex secretion of gustatory origin are responsible for the cephalic phase of gastric secretion. The arrival of food in the stomach in sufficient amount to distend it can act as a further stimulus to the gastric glands, but undoubtedly the humoral mechanism is the most important one in the gastric phase. Certain foods provoke the formation of gastrin by the mucosa of the pylorus; gastrin is absorbed and taken by the blood stream to the secretory cells of the stomach, which are thus stimulated into activity. This mechanism causes the persistence of secretion for a long time after the effects of the stimuli acting during the cephalic phase have ceased.

When food reaches the small intestine, gastric secretion is again stimulated by a humoral mechanism. It is yet doubtful whether this intestinal phase is due to the absorption of the products of digestion or to the formation in the duodenum and jejunum of a hormone that stimulates gastric secretion.

Inhibitory stimuli can act during any of the three phases. Certain psychic states, such as rage and fear, can stop or diminish gastric flow provoked by the stimuli mentioned above. Enterogastrone produced in the duodenum or jejunum also inhibits gastric secretion.

EXPERIMENTAL PEPTIC ULCERS

There are still many theories that endeavor to explain the causes of gastric and duodenal ulcers, but there is already a consensus as to which are the principal factors. Three main causes seem to have importance: (*a*) trauma to the mucosa, plus the corrosive action of hydrochloric acid and pepsin; (*b*) local circulatory disturbances of nervous or humoral origin; (*c*) chronic inflammation of the mucosa (gastritis).

The action of pepsin and hydrochloric acid seems

to be a factor common to all cases. Its importance is well demonstrated in Mann and Williamson's[1] procedure for the production of experimental ulcers: The duodenum is isolated by separating it from the pylorus and the jejunum. The proximal extremity of the duodenum is closed, and its distal end is joined to one of the end loops of the ileum. A gastrojejunostomy completes the operation (Fig. 174). By this means the bile and the pancreatic juice, which alkalinize the intestinal contents, are poured into the end of the small intestine, so that the acid chyme coming from the stomach is not neutralized. Chronic ulcers of the jejunum, situated close to the stomach opening, are observed in 95 per cent of animals so treated.

There are a great number of surgical techniques for the production of ulcers; they all combine trauma with the action of acid and pepsin.[2] Experimental ulcers can also be produced by prolonged treatment with atophan,[3] caffeine,[4] and histamine.

Vagotomy and peptic ulcer. Patients suffering from peptic ulcers secrete large quantities of gastric juice of high acidity, especially in the intervals between meals. Psychic factors acting through the vagi seem to play a major part in provoking this secretion. Double vagotomy at the level of the diaphragm has been performed in several of these cases, and a transitory decrease in the amount and acidity of gastric juice and in the movements of the stomach was observed.

[1] MANN, F. C., and C. S. WILLIAMSON, *Ann. Surg.*, **77**, 409, 1923.

[2] MARKOWITZ, J., "Textbook of Experimental Surgery," Wood, Baltimore, 1937.

[3] VAN WAGONER, F. H., and T. P. CHURCHILL, *Arch. Path.*, **14**, 860, 1932.

[4] JUDD, E. S. J., *Bull. Am. Coll. Surgeons*, **28**, 48, 1943; ROTH, J. A., and A. C. IVY, *Surgery*, **17**, 644, 1945.

Pain, gastric hypermotility, and nocturnal secretion of gastric juice were permanently decreased.[1]

BIBLIOGRAPHY

ALVAREZ, W. C., "An Introduction to Gastroenterology," 4th ed., Hoeber, New York, 1948.

BABKIN, B. P., The Digestive Work of the Stomach, *Physiol. Rev.*, **8**, 365, 1928.

BEAUMONT, W., "Experiments and Observations on the Gastric Juice, and the Physiology of Digestion," Plattsburgh, 1833.

BICKEL, A., Der Nervöse Mechanismus der Sekretion der Magendrüsen und der Muskelbewegung am Magendarmkanal, *Ergebn. d. Physiol.*, **24**, 228, 1925.

CANNON, W. B., "The Mechanical Factors of Digestion," E. Arnold and Co., London, 1911.

CARLSON, A. J., "The Control of Hunger in Health and Disease," University of Chicago Press, Chicago, 1919.

CARLSON, A. J., The Secretion of Gastric Juice in Health and Disease, *Physiol. Rev.*, **3**, 1, 1923.

CODE, C. F., The Inhibition of Gastric Secretion, *Pharmacol. Rev.*, **3**, 59, 1951.

IVY, A. C., The Mechanism of Gastric Secretion, *Surgery*, **10**, 861, 1941.

KLEE, P., Die Magenbewegung, in Bethe's "Handbuch der normalen und pathologischen Physiologie," **3**, 402, Springer, Berlin, 1927.

MCSWINEY, B. A., Innervation of the Stomach, *Physiol. Rev.*, **11**, 478, 1931.

MICHAELIS, L., Some Problems Concerning the Gastric Juice, *Harvey Lect.*, 1928, p. 59.

RYLE, J. A., "Gastric Function in Health and Disease," Oxford, New York, 1926.

WOLF, S., and H. G. WOLFF, "Human Gastric Function. An Experimental Study of a Man and His Stomach," Oxford, New York, 1943.

[1] DRAGSTEDT, L. R., and F. M. OWENS, *Proc. Soc. Exper. Biol. & Med.*, **53**, 152, 1943.

CHAPTER 38

Intestinal Digestion

FOOD REMAINS IN the stomach for a longer or shorter time, according to the nature of the substances ingested, the way they have been prepared and masticated, the motility of the stomach, etc. While the food is in the stomach it is attacked by saliva (which continues to act as long as the acidity is not too great), by the gastric enzymes, and by hydrochloric acid. The gastric juice and the movements of the stomach convert the gastric contents into a semiliquid paste, definitely acid, which has been called the chyme. This chyme is passed through the pylorus into the duodenum at intervals and in small quantities. In the intestine it comes into contact with the secretions of the pancreas, the intestine, and the liver, which will be considered separately.

PANCREATIC SECRETION

The pancreas is a gland that has an external and an internal secretion. The islets of Langerhans secrete insulin, which passes directly into the blood, while the cells of the acini secrete a fluid known as the pancreatic juice.

Methods of collecting pancreatic juice. Pancreatic juice can be obtained from temporary or permanent experimental fistulas. In the first case, Wirsung's duct is opened and catheterized; in the second case, the part of the duodenum where Wirsung's duct ends is sewn onto the skin.

Dogs with a permanent pancreatic fistula, in which the greater part of the pancreatic juice is lost, show serious disturbances (loss of appetite, asthenia, dehydration) after a few days and eventually die. Life can be prolonged by giving the animals a milk diet and intravenous injections of sodium bicarbonate and other electrolytes (Ringer's fluid), so as to replace minerals lost through the fistula. The administration by mouth of the pancreatic juice that flows from the fistula is especially efficacious in prolonging life. A better method consists in making a duodenal fistula, through which the pancreatic duct can be catheterized; there is no loss of pancreatic juice with this technique.[1]

In man several cases of pancreatic fistula, following a surgical operation or abdominal wounds, have been studied. The results have confirmed experimental observations on dogs.

Pancreatic juice can be collected by passing a double catheter with one opening in the stomach and the other in the duodenum; contamination with gastric contents can thus be prevented.[2]

The properties and composition of pancreatic juice. Pure pancreatic juice, obtained from a temporary or permanent fistula, is a colorless fluid, somewhat viscous, transparent or slightly opalescent, and alkaline (pH 8.4 to 8.9) because of the presence of sodium bicarbonate. It has proteins that are coagulated by heat and precipitated by alcohol. Pancreatic juice obtained by stimulation of the vagus nerve or injection of pilocarpine is more viscous than that which results from the injection of secretin (see page 364); it also has a greater enzymatic activity, and is less alkaline.

THE ACTIVITY OF PANCREATIC JUICE

Action on proteins. Since 1834 it has been known that pancreatic extracts attack proteins. Claude Bernard,[3] in 1856, obtained pure pancreatic juice by catheterizing Wirsung's duct and saw it had no effect on proteins when it was pure.

[1] THOMAS, J. E., and J. O. CRIDER, *Am. J. Physiol.*, **131**, 349, 1940.

[2] AGREN, G., and H. LAGERLÖF, *Acta med. Scandinav.*, **90**, 1, 1936.

[3] BERNARD, C., "Mémoires sur le pancréas," Paris, 1856.

Kühne[1] gave the name of "trypsin" to the substance responsible for the proteolytic effect of pancreatic juice. Shepovalnikov,[2] one of Pavlov's pupils, discovered the fact that intestinal juice activated the proteolytic effect of pancreatic juice; this was attributed by Pavlov to a substance that he called "enterokinase." Today it is known that pure pancreatic juice, collected by catheterization of Wirsung's duct, is inactive; it contains only a precursor or zymogen called "trypsinogen," which is transformed by a process of hydrolysis, activated by enterokinase (acting as an enzyme), into trypsin, a proteolytic enzyme. Trypsinogen and trypsin have been prepared in the crystalline state.[3]

Trypsin catalyzes hydrolysis of the peptide bonds of proteins and partially hydrolyzed proteins. Trypsin, in contrast to pepsin, attacks preferentially denatured proteins. For example, it has a weak action on natural egg albumen, seroglobulins, hemoglobin, collagen, etc., but digests these proteins easily if they have been previously modified by heat or other agents. Proteins are transformed by the action of trypsin into proteoses, peptones, polypeptides, and amino acids. Trypsin thus goes further than pepsin, which does not hydrolyze proteins as far as amino acids; nevertheless some polypeptides are not disintegrated even when the action of trypsin is considerably prolonged.

Trypsin develops its highest activity in an alkaline medium; the optimum pH varies between 8 and 9.7 for different substrates, but in a medium of such high alkalinity it is rapidly destroyed by autolysis. The duodenal contents are slightly acid (pH 4.5 to 5.1), because the acid chyme coming from the stomach mixing with the alkaline juices secreted by the pancreas, the duodenum, and the liver is only partially neutralized. Trypsinogen is rapidly transformed into trypsin by the enterokinase in the duodenum. The optimum pH for the enzymatic effect of enterokinase is 5.2 to 6. Trypsinogen is also transformed into trypsin by autocatalysis[4] at pH 7 to 9.

Action on milk. Another zymogen in the pancreatic juice is chymotrypsinogen, which is converted by trypsin into chymotrypsin in the duodenum. This enzyme is not trypsin, and enterokinase does not activate its zymogen. Both the enzyme and the zymogen have been obtained in crystalline form. This enzyme has a powerful curdling effect on milk.

[1] Kühne, W., *Virchow's Archiv*, **39**, 130, 1867.

[2] Shepovalnikov, N. P., The Physiology of Intestinal Juice, thesis, St. Petersburg, 1899 (in Russian).

[3] Northrop, J. H., and M. Kunitz, *J. Gen. Physiol.*, **16**, 267, 1932–1933; Kunitz, M., and J. H. Northrop, *Science*, **80**, 505, 1934.

[4] Kunitz and Northrop, *op. cit.*, p. 190.

Action on carbohydrates. Pancreatic juice contains an enzyme that acts like ptyalin, the amylase of saliva, hydrolyzing starch down to maltose. This amylase (amylopsin) has an optimum pH of 6.7 to 7.2. A maltase, which catalyzes the hydrolysis of maltose into glucose, is also found in pancreatic juice.

Action on fats. Claude Bernard demonstrated in a series of experiments, now classic, the emulsifying and hydrolytic action of pancreatic juice on fats.[1] A powerful esterase, known as the "pancreatic lipase" (steapsin), hydrolyzes neutral fats into fatty acids and glycerol. The hydrolysis of other esters is also catalyzed by this enzyme. Its optimum pH is about 8.0.

In an alkaline medium, bile has an important activating effect on pancreatic lipase. This effect is due exclusively to the bile salts. The absence of pancreatic juice causes serious disturbances in digestion, especially in the digestion and absorption of fats and fat-soluble vitamins.

Bicarbonate in pancreatic juice neutralizes the acidity of gastric juice and regulates the pH not only of the intestinal contents but also of the stomach when the contents of the duodenum are regurgitated into the stomach. Gastric function may thus be disturbed if there is no secretion of pancreatic juice.

The Secretion of Pancreatic Juice

Pancreatic juice is secreted intermittently in some animals; in others, *e.g.*, in the dog, rabbit, rat, guinea pig, and perhaps in man, it is secreted continuously. This continuous or "spontaneous" secretion may be caused by certain nervous and hormonal factors. The ingestion of food is followed after a short time by abundant flow of pancreatic juice.

Nervous mechanism. The pancreas is innervated by sympathetic and parasympathetic (vagus) fibers. Electric stimulation of the peripheral end of the cut vagus produces secretion of a small amount of pancreatic juice, rich in enzymes; there is an important ecbolic and a moderate hydrelatic effect. Secretin and vagal impulses act synergically; stimulation of the

[1] Bernard, *op. cit.*

vagus increases the response to secretin, and reciprocally the previous injection of secretin increases the response to stimulation of the vagus. Atropine suppresses the effects of vagal stimulation. Pilocarpine, acetyl-β-methylcholine, and other parasympathomimetic drugs have an ecbolic effect on the pancreas. Vagal stimulation can also interrupt or diminish the secretion provoked by other stimuli. This apparent secretory inhibition is probably due to the stimulation of fibers contained in the vagus nerve, which provokes contraction of the pancreatic ducts and prevents the flow of juice, which is thus retained within the ducts.

Stimulation of the sympathetic nerves also provokes the secretion of pancreatic juice rich in organic substances and enzymes. This effect is potentiated by eserine and inhibited by atropine; therefore it seems to be the result of the activity of cholinergic fibers in the sympathetic. Adrenaline and other sympathomimetic drugs have a predominantly inhibitory effect on pancreatic secretion, mainly due to vasoconstriction.

Humoral mechanism. Bayliss and Starling[1] showed that the introduction of acid into a loop of the jejunum, after severing all its nervous connections with the rest of the organism, provoked a flow of pancreatic juice in all respects the same as that observed in normal animals. It was already known that injection of acids into the blood stream did not stimulate pancreatic secretion.

The only barriers between the intestinal cavity and the blood stream are the intestinal mucosa and the capillary endothelium. Bayliss and Starling thought that acid must act on the intestinal epithelium, extracting a substance which, passing into the blood stream, stimulated the pancreatic cells. They prepared an extract of duodenal mucosa in 0.4 per cent HCl and, after neutralizing it, injected it intravenously; they obtained an intense secretory response. This now classic experiment demonstrated for the first time the existence in the organism of chemical substances that can stimulate distant organs into activity. These chemical messengers were called "hormones" by Bayliss and Starling.[2]

Conclusive evidence that the stimulus for pancreatic secretion is transmitted by the blood

[1] Bayliss, W. M., and E. H. Starling, *J. Physiol.*, 28, 325, 1902.

[2] See Chap. 51.

from the duodenum to the pancreas was given by experiments in which a duodeno-pancreas was grafted into the neck of a normal dog.[1] Injection of acid into the duodenum of the dog provoked a flow of juice from both the pancreas *in situ* and the grafted pancreas; therefore a chemical messenger goes in the blood from the duodenum to the pancreas and stimulates its secretory activity (Fig. 175).

The hormone in this mechanism was called "secretin" by Bayliss and Starling. It is found in the duodenal mucosa, and can be extracted not only by acids, but also by many other substances, such as water, soapy solutions, weak alkalis, etc. The amount of secretin that can be extracted from the intestinal mucosa diminishes as the distance from the pylorus increases.

Properties of secretin. Secretin has been purified, and crystals of its picrolonate have been obtained.[2] Apparently it is a polypeptide of low molecular weight (5,000), easily soluble, and diffusible through colloid membranes; it is destroyed by pepsin and trypsin. Its principal action is on the pancreas, but it also stimulates Brünner's glands of the duodenum, and the secretion of intestinal juice and bile. Secretion increases the flow of, and also the concentration of, bicarbonate in the pancreatic juice, but this juice is poor in enzymes, contrasting thus with the juice secreted in response to stimulation of the vagus nerve, or injection of choline derivates (mecholyl, etc.). Ergotamine and dehydroergotamine suppress the pancreatic response to secretin.[3]

Another hormone has been found in the duodenal mucosa; it is called "pancreozymin"[4] and it increases the enzymatic concentration of pancreatic juice, especially that of amylase. Pancreozymin can be separated from secretin by means of aniline, and it loses its activity after prolonged incubation with blood serum.[5]

Adaptation of the secretion of the principal enzymes to the diet has been demonstrated. Thus, after a time, a carbohydrate diet produces

[1] Delezenne, C., L. Hallion, and R. Gayet, *Ann. physiol. physicochim. biol.*, 3, 508, 1927; Houssay, B. A., and E. A. Molinelli, *Rev. Soc. argent. de biol.*, 3, 362, 1927.

[2] Hammarsten, E., G. Argen, H. Hammarsten, and O. Wilander, *Biochem. Ztschr.*, 264, 275, 1933.

[3] Greengard, H., C. D. Collins, and A. C. Ivy, *Federation Proc.*, 4, 26, 1945.

[4] Harper, A. A., and H. S. Raper, *J. Physiol.*, 102, 115, 1943.

[5] Greengard, H., M. I. Grossman, J. R. Woolley, and A. C. Ivy, *Science*, 62, 350, 1944.

pancreatic juice with a high concentration in amylase and a protein diet produces a juice rich in trypsin. On the other hand, fatty diets do not increase the content of lipase.[1] This may be the result of a process of adaptation, in the sense that ingestion of a certain kind of food stimulates production and secretion of the corresponding enzymes; or it may be that certain foodstuffs are necessary for the synthesis of enzymes, which does not take place if these substances are not found in the diet.[2] According to Babkin's observations in dogs and Lagerlöf's in men, the concentration of the three principal enzymes in pancreatic juice varies simultaneously, *e.g.*, when protease increases, lipase and amylase also increase.

Regulation of Pancreatic Secretion

There are a cephalic phase and an intestinal phase in the course of pancreatic secretion. There is no gastric phase; *i.e.*, introduction of food into, or distention of, the stomach does not provoke pancreatic secretion.[1]

The cephalic phase. The introduction of food into the mouth starts a reflex which produces the secretion of a small amount of pancreatic juice rich in enzymes. Atropine and double vagotomy suppress this reflex.

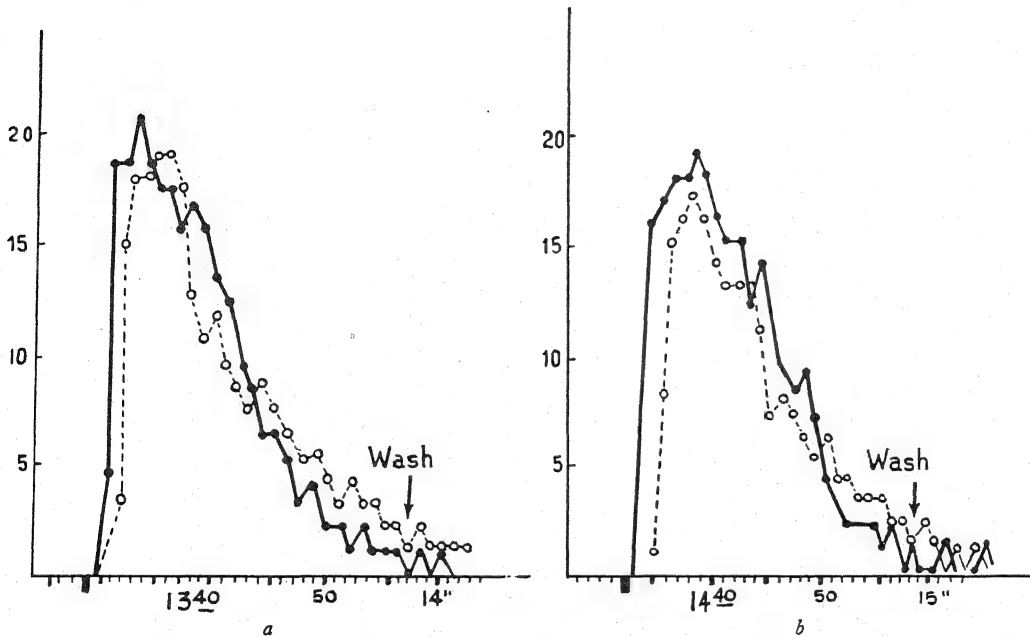



Fig. 175. Comparison of the effects produced on the secretions of the pancreas *in situ* and of the pancreas grafted into the neck, as shown by the number of drops of juice secreted per minute. Hydrochloric acid (30 cc., 0.5 per cent solution) was introduced into the small intestine *in situ* (*a*) and into a loop of small intestine grafted into the neck (*b*). Solid line: secretion of the grafted pancreas; broken line, secretion of the pancreas *in situ*. (*Gayet, R. "Titres et travaux scientifiques," Masson et Cie, Paris, 1933.*)

The intestinal phase. This is the most important stage of pancreatic secretion. All the principal constituents of the chyme (water, HCl, products of protein digestion, fats, fatty acids, digestive derivates of starch) stimulate pancreatic secretion when they are introduced into the duodenum. These substances produce the release of secretin, but the juice secreted in response to the introduction of foodstuffs into the duodenum has a higher concentration of enzymes than the juice obtained by injection of secretin or the introduction of HCl into the duodenum. The intestinal phase is, therefore, not due exclusively to secretin, but also to other factors, hormonal (pancreozymin) or nervous (true reflexes or local reflexes). Thus,

[1] Grossman, M. I., H. Greengard, and A. C. Ivy, *Am. J. Physiol.*, **138**, 676, 1943.

[2] Thomas, J. E., "The External Secretion of the Pancreas," Charles C Thomas, Springfield, Ill., 1950.

[1] Ivy, A. C., *Ann. Clin. Med.*, **4**, 798, 1925.

a grafted pancreas responds to the introduction of foodstuffs into the intestine with an increase in concentration of enzymes secreted in the juice, *i.e.*, a response corresponding to the release of pancreozymin.[1]

The formation of sodium bicarbonate. Probably enzymes are secreted by the cells of the acini, and the cells of the intralobular ducts secrete water and bicarbonate. The pancreas of an animal intoxicated with alloxan secretes enzymes normally, but the amount of fluid diminishes. This drug damages intralobular duct cells but does not affect the cells of the acini.[2] The pancreatic acini have the appearance of inverted gastric tubular glands. This fact, together with the results of alloxan intoxication, has led Thomas[3] to formulate the hypothesis that intralobular duct cells, and perhaps centroacinar cells, form bicarbonate by a mechanism similar to that by which HCl is formed in the glands of the stomach.

INTESTINAL SECRETION

The secretion of the glands of the mucosa of the small intestine is called intestinal juice. The crypts or glands of Lieberkühn secrete the intestinal juice; Brünner's glands, found exclusively in the first part of the duodenum, secrete an alkaline juice which is rich in mucus and which contains enterokinase, amylase, and traces of esterase.[1]

Methods of collecting intestinal juice. In the experimental animal the best way of obtaining intestinal juice and of studying its secretion is to make a permanent intestinal fistula. The Thiry and Thiry-Vella fistulas consist in the separation of a loop of intestine, with undisturbed innervation and circulation, and the connection of one or both ends of the loop to the skin (Fig. 176*a* and *b*). Pure intestinal juice can thus be obtained. To collect the intestinal contents metallic cannulas are placed at different levels in the gut (Pavlov). Mann and Bollman have substituted for the cannulas a strip of intestine (Fig. 176*c*).

In man accidental and surgical intestinal fistulas have given the opportunity of studying intestinal digestion. The intestinal juice has also been collected by introducing a rubber catheter into the gut through the mouth and stomach.

The properties and composition of enteric juice. Intestinal juice is usually a cloudy fluid, due to the presence of leukocytes, epithelial cells, and mucus. Sodium bicarbonate makes it more or less alkaline (pH 7 to 8.6). Apart from enterokinase, which transforms trypsinogen into trypsin, it contains several enzymes. The enzymes are found preformed in the mucosa, the extracts of which are more potent than enteric juice. Certain authors[2] maintain that only amylase and

a *b* *c*

Fig. 176. Different types of intestinal fistula. *a*, Thiry fistula; *b*, Thiry-Vella fistula; *c*, Mann-Bollman fistula. (*Markowitz, "Textbook of Experimental Surgery," William Wood & Combanv, Baltimore, 1937.*)

[1] Wang, C. C., and M. I. Grossman, *Am. J. Physiol.*, **164**, 527, 1951.

[2] Grossman, M. I., and A. C. Ivy, *Proc. Soc. Exper. Biol. & Med.*, **63**, 62, 1946.

[3] Thomas, J. E., *Tr. New York Acad. Sc.*, **14**, 310, 1952.

[1] Florey, H. W., R. D. Wright, and M. A. Jennings, *Physiol. Rev.*, **21**, 36, 1941.

[2] Wright, R. D., M. A. Jennings, H. W. Florey, and R. Lium, *Quart. J. Exper. Physiol.*, **30**, 73, 1940.

enterokinase are truly secreted by the cells of the intestinal mucosa. They hold that other enzymes (peptidase, invertase, lipase, etc.) found in the lumen of the gut are not secreted but form part of the cells shed by the mucosa. Others[1] state that the four main enzymes of enteric juice are secreted simultaneously.

The amount and composition of enteric juice differ according to the level of the gut at which it has been collected and the nature of the stimulus used. As the distance to the pylorus increases, the amount of juice secreted and its enzymatic activity diminish.

The first part of the duodenum, where the Brünner glands are found, secretes an alkaline juice with abundant mucus. This secretion is stimulated by the local action of acids or foodstuffs, which release a hormone, *duocrinin*,[2] different from secretin.

The Activity of Enteric Juice

Action on peptides. Cohnheim gave the name of "erepsin" to the enzyme in the intestinal mucosa that acts on peptides. This is really an enzyme complex (aminopolypeptidase, dipeptidase, carboxypolypeptidase, polypeptidase)[3] which attacks the products of partial disintegration of proteins and transforms them into amino acids. Enteric juice does not act on native proteins, nor on proteoses and peptones.

Intestinal juice also contains *nucleinases*, which disintegrate nucleic acids into nucleotides, *nucleotidases*, which transform these nucleotides into phosphoric acid and nucleosides, and *nucleosidases*, which split nucleosides into pentose and purine bases. *Arginase* and *phosphatase* are found in the intestinal mucosa.

Action on fats. The intestinal juice has a weak *lipase*, which nevertheless can hydrolyze a good part of the fats in the food when there is no pancreatic juice.

Action on carbohydrates. The enteric juice completes the disintegration of carbohydrates which commences in the mouth and which is continued in the stomach by the saliva swallowed with the food and in the duodenum by the pancreatic amylase. The enzymes in saliva and pancreatic juice give disaccharides as final products; the enteric juice completes hydrolysis down to monosaccharides, of which only small quantities have been formed by the action of the HCl in the gastric juice. The intestinal juice has a *saccharase* or *invertase*, which disintegrates sucrose into fructose and glucose; a *maltase*, which transforms maltose into glucose; and in young mammals a *lactase*, which converts lactose into galactose and glucose.

[1] Schiffrin, M. J., and E. S. Nasset, *Am. J. Physiol.*, **128**, 70, 1939.

[2] Grossman, M. I., *Physiol. Rev.*, **30**, 33, 1950.

[3] Aminopolypeptidase is an enzyme acting on polypeptides that have a basic N with at least one H. Carboxypeptidase, or carboxypolypeptidase, acts on polypeptides that have a free carboxyl but no free amines. Polypeptidases hydrolyze polypeptides; they act on the peptide bond next to the free amine. Dipeptidase disintegrates dipeptides made up of natural amino acids.

The Secretion of Intestinal Juice

There is almost general agreement that no secretion flows from an intestinal fistula in response to tasting and swallowing food. On the other hand, mechanical stimulation of the fistula provokes an abundant secretion. Local electrical stimulation and the local application of certain chemical substances also provoke secretion. Local stimulation is efficacious even after the intestine has been denervated; for this reason it is supposed that local stimulation produces its effects through reflexes taking place in Meissner's intramural plexus, or by a hormonal mechanism.

Vagal stimulation seems to produce enteric secretion.[1] Sympathectomy is followed by an abundant flow of "paralytic" secretion. In normal conditions, therefore, the sympathetic probably inhibits intestinal secretion.

Parenteral administration of several drugs (histamine, pilocarpine, eserine, and secretin) is followed by secretion of intestinal juice.

A substance, which is not secretin, has been extracted from the intestinal mucosa. The injection of this substance, called "enterocrinin," provokes an abundant secretion of enteric juice.[2] Conclusive evidence that enterocrinin is a specific hormone, which regulates intestinal secretion, has not yet been obtained. Concentrated preparations of enterocrinin with high potency have been obtained.[3]

In normal conditions the main stimulus of intestinal secretion is the local effect of mechanical stimulation and chemical agents found in the chyme.

[1] Savich, V. V., and N. A. Shosestvensky, *Compt. rend. Soc. de biol.*, Paris, **80**, 508, 1917.

[2] Nasset, E. S., *Am. J. Physiol.*, **121**, 481, 1938.

[3] Fink, R. M., *Am. J. Physiol.*, **139**, 633, 1943.

THE BILE

The bile can be considered as both an excretion and a secretion. It is an excretion because it serves to eliminate the bile pigments, which result from the disintegration of hemoglobin, and some organic or mineral substances, together with certain drugs used therapeutically. It is a secretion because of its important digestive function. The liver produces bile continuously, but the presence of Oddi's sphincter and the gall bladder allows it to be poured intermittently into the intestine. Some animals, *e.g.*, the rat and the horse, do not have a gall bladder; this lack is compensated, in part, by the greater capacity of the bile ducts. The walls of the gall bladder, and in a lesser degree those of the ducts, secrete mucus and absorb water, so that storage of the bile results in its concentration.

Methods of study. Several surgical methods for permanent fistulization of the bile ducts have been devised. A catheter can be introduced into the common bile duct, or the area of the duodenal mucosa containing the ending of this duct (Vater's ampulla) can be transplanted to the skin. In man a temporary bile fistula, made with the purpose of evacuating the bile, is frequently used in the treatment of biliary disease. Duodenal catheterization by mouth is also commonly used to collect the duodenal contents.

Physical Properties of Bile

The color of the bile varies with its concentration and the pigments it contains. Hepatic bile is golden yellow or orange colored in man. The bile in the gall bladder is darker. The specific gravity of hepatic bile is approximately 1.008; that of the bile in the gall bladder can be up to 1.050. The osmotic pressure of hepatic bile is approximately that of blood (freezing point, −0.56 to −0.61°C.). Hepatic bile is slightly alkaline (pH 7.3 to 7.7) and that of the gall bladder is usually slightly acid (pH 6 to 7).

Chemical Composition of Bile

The bile obtained directly from the liver is less concentrated than gall-bladder bile. The principal components of bile are mucin, bile salts, bile pigments, lecithin, and cholesterol. Bile salts are the only substances of importance in digestion.

The bile salts. These are complex mixtures of the sodium salts of taurocholic and glycocholic acids. These acids are formed by combination of taurine and glycine respectively with one of the cholic acids (cholic, desoxycholic, and lithocholic acids). Cholic acid has a chemical structure similar to that of cholesterol, and it has recently been demonstrated that it is formed from cholesterol in the liver.[1] Taurine is formed from cystine. Glycine is an amino acid that can be synthesized by the organism.

Table 32. Composition of Human Bile

Constituent	*Hepatic bile, %*	*Vesicular bile, %*
Water	97.48	83.98
Mucin and pigments	0.53	4.44
Bile salts	0.93	8.70
Fatty acids (soaps)	0.12	0.85
Cholesterol	0.06	0.87
Lecithin	0.02	0.14
Mineral salts	0.83	1.02

Source: Hammarsten.

Bile salts lower the surface tension of solutions and thus further the emulsification of fats in the duodenum. The salts of desoxycholic acid also help dissolve many substances normally insoluble in water (fatty acids, lecithin, cholesterol, etc.). This "hydrotropic" effect is of great importance in the digestion and absorption of fats.

The bile salts excreted into the duodenum are reabsorbed and carried by the portal blood back to the liver, whence they are again excreted. Bile salts given by mouth can be almost completely recovered in the bile; they are therefore totally reabsorbed. This enterohepatic circulation of bile salts represents a great economy of effort for the organism. Nevertheless when the bile is lost through a biliary fistula, the liver continues to produce bile salts. Nothing is yet known about the factors that regulate the production of bile salts.

Bile pigments. These substances are responsible for the color of the bile. The golden yellow color in man is given by bilirubin, the principal pigment, and the green color by biliverdin. Biliverdin is a product of oxidation of bilirubin, and it can be found in vesicular bile. In the toad, biliverdin is the only pigment in the bile, so the accumulation of bile pigments in the blood following the obstruction of the common bile duct

[1] Bloch, K., B. N. Berg, D. Rittenberg, *J. Biol. Chem.*, **149**, 511, 1943.

causes the tissues to take on a greenish color instead of the yellow color typical of jaundice in mammals.[1]

The bile pigments are formed in the reticuloendothelial cells of the liver and other parts of the body in the process of disintegration of hemoglobin. Injection of hemoglobin or an excessive destruction of red blood cells causes an increase in the formation of bile pigments, which pass into the blood and are then excreted by the liver into the bile.

Bilirubin is reduced in the intestine by the action of bacteria, and is converted into urobilinogen, which is in part eliminated in the feces as stercobilinogen. A smaller amount of urobilinogen is reabsorbed and fixed by the liver, whence it goes to the blood and is excreted by the kidney. Stercobilinogen and urobilinogen are easily oxidized (in the gut or the renal excretory tubes), and thus stercobilin and urobilin are produced (see Chap. 4).

Normally up to 0.6 mg. of urobilinogen and urobilin are eliminated daily. The excretion of larger quantities is a sign of liver insufficiency, as it shows that there is a diminished capacity of the liver to retain this substance after reabsorption from the gut.

Cholesterol. Cholesterol is more readily soluble in bile than in water, because of the action of bile salts. It appears in high concentration in bile. Cholesterol is excreted with the bile into the duodenum, where it facilitates emulsification and absorption of fats and other lipids. Part of it is reabsorbed, but part is transformed into coprosterol by the intestinal bacteria and is eliminated in the feces.

The Action of Bile in Digestion

Action on fats. The favorable effect of bile in the digestion and absorption of fats is due to the bile salts and is of considerable importance. The activity of pancreatic lipase is increased several times by the presence of bile. This effect is due to the lowering of the surface tension by bile salts, which furthers the emulsification of fats; it is also due to the hydrotropic effect of bile salts, which increases the solubility of the fatty acids and soaps (even calcium salts, which are insoluble in water) that are produced by the digestion of fats. Bile also increases the effect of pancreatic lipase on dissolved substances, perhaps because in some way it helps the union of enzyme and substrate. The absorption of soaps and fatty acids is favored by the solvent action of bile. Cholesterol, the fat-soluble vitamins D, K, and E, and carotene need the presence of bile in the intestine in order to be absorbed.

An antiseptic or bacteriostatic action on the intestinal flora has been attributed to bile.

Action on intestinal motility. Obstruction of the biliary ducts is accompanied by a diminished intestinal motility. This might be due either to the lack of bile in the intestine or to the retention of bile salts or other bile constituents.[1] The administration of bile increases intestinal motility.

The Secretion of Bile

The secretion of bile by the liver is a continuous process, and it must be distinguished from the evacuation of bile into the intestine, which takes place intermittently.

The continuous secretion of bile is not dependent on nervous activity, as it is not suppressed by denervation. Nevertheless stimulation of the vagus increases and stimulation of the sympathetic nerves inhibits bile secretion.[2] Fasting and the loss of bile through a fistula diminish the secretion of bile. If a manometer is placed in the common bile duct, the pressure gradually rises to 250 to 300 mm. of bile and remains there. Secretion ceases at this pressure and jaundice appears. In normal conditions when the pressure reaches 50 to 70 mm. the bile flows through the cystic duct into the gall bladder. A normal adult man secretes 15 cc. of bile per kilogram of body weight every 24 hr.—about 1,000 to 1,100 cc. in a man weighing 60 to 73 kg.

The amount of bile secreted by the liver is increased by several substances given by mouth or injected into the blood stream. These substances are called choleretics.[3] Bile salts and bile acids are the most potent of all choleretics. The administration of 10 gm. of taurocholic acid, by mouth or injection, is followed by a rise in bile

[1] Cabello, J., *Rev. Soc. argent. de biol.*, **19**, 81, 1943.

[1] Canonico, A., and F. C. Mann, *Surgery*, **13**, 81, 1943.

[2] Tanturi, C. H., and A. C. Ivy, *Am. J. Physiol.*, **121**, 270, 1938.

[3] The term "cholagogue" (Greek, χολή, bile, and ἄγειν, to conduct) has long been used to signify an agent that increases the flow of bile by an increase either in the bile secreted or in that evacuated. More recently its significance has been limited to a greater evacuation of bile, and the term "choleresis" (Greek χολή, bile, and ερησις, removal) serves to indicate an increase in the bile secreted by the liver.

secretion that lasts over 24 hr. Secretin and stimulation of the vagus nerve also increase bile secretion. Among foodstuffs the most effective choleretics are fats, fatty acids and soaps, and the products of protein digestion. Atophan, eserine, and pilocarpine are also choleretics.

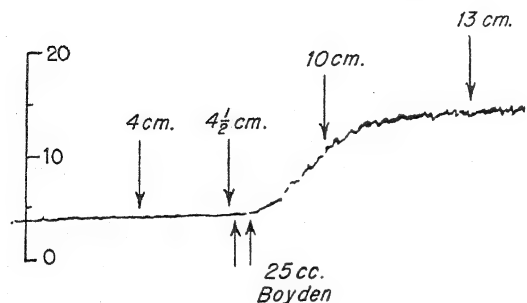



FIG. 177. Pressure in the common bile duct. On the left, pressure scale in cm. H_2O. Between the arrows, 25 cc. of milk, cream, and egg yolk (Boyden's meal) is introduced into the duodenum. Lower tracing, time in minutes. (*After Royer, M., and F. J. Manfredi, Rev. Soc. argent. de biol., vol. 20, p. 232, 1944.*)

Bile secretion is diminished by sugar in the diet, by fasting, by the stimulation of the sympathetic nerves, by the injection of adrenaline or morphine, and by anesthetics.

STORAGE AND EVACUATION OF BILE

The liver pours the bile into the intrahepatic ducts, which end in the hepatic ducts; these join to form the common bile duct. The cystic duct, which comes from the gall bladder, also ends in the common bile duct. The gall bladder is a reservoir in which bile is accumulated, modified, and concentrated during the intervals between digestive periods. The common bile duct ends next to Wirsung's pancreatic duct in Vater's ampulla in the duodenum. Royer and Solari[1] have shown that a dilatation of the end of the common bile duct, forming Vater's ampulla, occurs only in exceptional cases. The radiographic study of the duct filled with an opaque substance shows it ending in a beveled edge. Sometimes Wirsung's duct has a separate opening into the duodenum. As it passes through the intestinal wall, the common bile duct is surrounded by smooth-muscle fibers, which can close the duct when they contract. The opening of the duct into the ampulla is surrounded by a ring of smooth-muscle fibers which form Oddi's sphincter.

The liver secretes bile continuously during the intervals between the digestive periods. In dogs in the basal state the pressure in the common bile duct is 20 to 40 mm. H_2O, with rhythmical oscillations.[1] When the gall bladder empties, the pressure in the duct increases rapidly (Fig. 177). The sphincter of Oddi is in a state of tonic contraction, which can resist pressure up to 300 mm. H_2O. This tonic contraction serves to prevent the flow of bile into the gut and the entrance of the acid chyme and intestinal bacteria into the bile duct.

When the pressure within the duct rises to between 50 and 70 mm. H_2O, bile flows through the cystic duct into the gall bladder.

The gall bladder is a reservoir that contains about 50 cc. in man. There the bile is concentrated by the absorption of salts and water, and mucus secreted by the cells of the mucosa is added to it. The bile in the gall bladder is about 10 times as concentrated as bile newly secreted by the liver; therefore the 50 cc. contained in the gall bladder is the equivalent of 500 cc. of hepatic bile.

The arrival of the chyme into the duodenum provokes relaxation of Oddi's sphincter, contraction of the gall bladder, and probably also contraction of the bile ducts. Bile is thus evacuated into the gut. Afterward Oddi's sphincter contracts, the gall bladder relaxes, and the evacuation of bile ceases. These periods alternate during the first hours of digestion.

Cholecystography and duodenal catheterization. There are two principal methods for exploring gall-bladder function in man: cholecystography (the Graham-Cole test) and duodenal catheterization (the Meltzer-Lyon test).

Cholecystography consists in the x-ray examination of the gall bladder. A substance that

[1] ROYER, M., and A. V. SOLARI, *Gastroenterology*, **2**, 180, 1944.

[1] ROYER, M., and F. J. MANFREDI, *Rev. Soc. argent. de biol.*, **20**, 232, 1944.

is eliminated by the liver and therefore passes into the gall bladder (tetraiodophenolphthalein or tetrabromophenolphthalein) is given by mouth or injected intravenously. As the bile is concentrated in the gall bladder the concentration of the substance, which is opaque to x-rays, also increases, so the gall bladder is visualized in the x-ray film. The shape, size, and position of the gall bladder can thus be established. The ingestion of emulsified fats (egg yolk, cream), and to a lesser degree that of protein, cause the evacuation of the gall bladder in 1½ hr., as can be seen by taking a series of x-ray films.

In abnormal cases the bladder cannot be visualized with this method for one of the following reasons: (*a*) insufficient elimination of the drug by the liver; (*b*) incontinence of Oddi's sphincter; (*c*) obstruction of the cystic duct or of the gall bladder; (*d*) insufficient concentration in the gall bladder.

Duodenal catheterization consists in having the subject swallow a rubber tube of adequate diameter and length and passing it into the duodenum, the contents of which are drawn off with a syringe.

A few cubic centimeters of a solution of 33 per cent magnesium sulfate, or of olive oil, injected into the gut through the catheter, provokes a flow of bile and the evacuation of the gall bladder. At first a yellow-colored bile (bile A) coming from the liver is obtained; then a dark green, viscous bile (bile B), coming from the gall bladder; and finally again yellow bile from the hepatic ducts. The absence of B bile has a significance similar to that of the absence of visualization of the gall bladder, *i.e.*, it means that this organ is not functioning normally.

The motility of the gall bladder. The gall bladder has two different types of contractions: (*a*) tonic contractions, which increase the pressure within it up to 300 mm. H_2O and last 10 to 30 min. or even longer; (*b*) rhythmic contractions at the rate of two to six per minute. The gall bladder has also a postural tonus, by which is meant the tendency to recover the resting pressure after it has been altered by any cause. Tonic contractions can appear spontaneously or as a consequence of nervous or chemical stimuli.

The vagus nerve and the sympathetic nerves have motor fibers going to the gall bladder; the sympathetic nerves also have inhibitory fibers. The vagus is predominantly the motor nerve and the sympathetic nerves are predominantly inhibitory. Several drugs act on the motility of the gall bladder;[1] sympathomimetic drugs (adrenaline) relax the gall bladder and parasympathomimetic drugs (acetylcholine, pilocarpine) contract it. Atropine suppresses the effect of the parasympathomimetic drugs. Cholecystokinin (see below) also provokes contraction of the gall bladder.

The mechanism of bile evacuation. The evacuation of bile into the duodenum is dependent on several factors. The most important are the contraction of the gall bladder, the condition of Oddi's sphincter and of the duodenal musculature, and choleresis.

The principal stimulus of bile evacuation is the ingestion of food. Fats are the most efficacious; proteins are also active, while carbohydrates are much less so. Among the fats, egg yolk is the most potent. Olive and other oils, cream, etc., also show activity.[2] Acid in the duodenum provokes evacuation of the gall bladder.

Boyden[3] has demonstrated the intermittent evacuation of bile in man. After a meal there are alternate periods of contraction and relaxation in the gall bladder. Three phases can be distinguished in the periods of contraction:

1. An initial response commences 1 min. after the ingestion of food and lasts 2 min. It may be due to relaxation of the sphincter and contraction of the gall bladder by a psychic reflex, or to the arrival of food to the stomach or duodenum.
2. A pause of 2 min. follows the initial response. It is probably due to the recovery of tonus by Oddi's sphincter.
3. The main discharge then takes place. It lasts 6 to 60 min., and half to two-thirds of the contents of the gall bladder is evacuated.

A period of rest follows, lasting 5 to 45 min. These alternating periods are repeated several times until the gall bladder is completely emptied.

In man morphine increases the tonus of Oddi's sphincter; magnesium sulfate is supposed to re-

[1] Houssay, B. A., and H. H. Rubio, *Rev. Soc. argent. de biol.*, **8**, 370 and 379, 1932.

[2] These substances must pass into the duodenum before they act. Sometimes they remain for a long time in the stomach, and therefore their action on the gall bladder is delayed.

[3] E. A. Boyden, *Anat. Rec.*, **40**, 147, 1928.

lax it. In the dog opiates diminish the resistance of Oddi's sphincter.[1]

A reflex mechanism for evacuating the gall bladder has also been suggested. Some workers describe a mechanism of reciprocal innervation by means of which, on the ingestion of food or the arrival of acids or salts into the duodenum, a reflex is started that produces relaxation of Oddi's sphincter and contraction of the gall bladder. Conclusive evidence for this statement is still to be obtained.

Definite proof has been given that the denervated gall bladder, or a grafted gall bladder, can contract and evacuate its contents in response to the introduction of acids or fats into the duodenum. Contraction of the gall bladder is due to a hormone, which is similar to, but not identical with, secretin.[2] Ivy and Oldberg[3] have named it "cholecystokinin." Acids and the products of the digestion of fats introduced into the duodenum form or release this hormone, which passes into the blood stream and stimulates the contraction of the gall bladder.

THE MOTILITY OF THE SMALL INTESTINE

Methods of study. Several methods have been used in the study of the movements of the intestine: (*a*) direct observation or cinematography of the intestines, either after opening the abdomen with the animal submerged in saline solution, or else through a window in the abdominal wall; (*b*) direct observation or graphic registration of the movements of an intestinal loop transplanted out of the abdomen and covered by a thin cuff of skin (Ivy); (*c*) observation with x-rays after ingestion of an opaque meal; (*d*) the introduction of balloons into the gut by mouth, by rectum, or through fistulas, for registration of changes in intra-intestinal pressure; (*e*) registration of the movements of isolated strips of intestine.

a

b

c

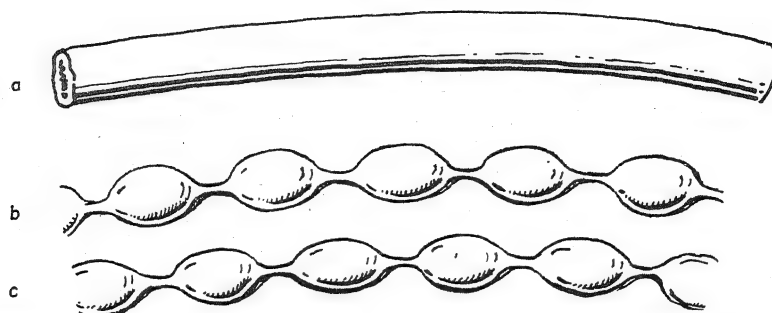



Fig. 178. Diagram of segmentation movements in the small intestine: *a*, at rest; *b*, first stage; *c*, second stage.

Movements of the Small Intestine

It is usual to describe three types of movements in the small intestine: (*a*) rhythmic movements of segmentation; (*b*) peristalsis; (*c*) pendular movements.

Rhythmic movements of segmentation. Cannon considers this type of movement as the most constant and fundamental form of intestinal motility. A quiescent loop of intestine suddenly shows a regular series of annular constrictions, which divide it into ovoid segments. Shortly afterward, a constriction appears in the most dilated part of each segment and simultaneously the first constrictions disappear (Fig. 178). This alternation is repeated six to ten times per minute in man. The intestinal contents do not progress; they remain within the loop. These movements serve (*a*) to mix the contents with the intestinal juices; (*b*) to bring the intestinal contents into contact with the villi; (*c*) to evacuate the lymphatics and veins of the intestinal wall and thus favor local circulation.

The movements described are not dependent on the intrinsic or extrinsic innervation of the gut; they can be observed after complete denervation, and also in strips of intestinal muscle in which the nerve plexuses have been removed. Nicotine and cocaine, drugs that paralyze nervous activity, do not modify these movements. Rhythmic segmentation, therefore, is apparently myogenic.

[1] Manfredi, F. J., and M. Royer, *Rev. Soc. argent. de biol.*, **22**, 67, 1946.

[2] Agren, G., *Skandinav. Arch. f. Physiol.*, **81**, 234, 1939.

[3] Ivy, A. C., and E. Oldberg, *Am. J. Physiol.*, **86**, 599, 1928.

According to Alvarez[1] the frequency and force of contractions, intestinal tone, and irritability diminish progressively from the duodenum to the ileocolic valve, coinciding with a gradual decrease in the metabolic activity (oxygen consumption and CO_2 production) of the intestinal muscle as the distance from the pylorus increases. This parallelism between muscular activity and metabolic exchanges led Alvarez to put forward his theory of the existence of a metabolic gradient in the intestine, on which its physiologic activity and peristaltic polarity would depend.

Peristaltic movements. These movements are the same as those of peristalsis already described in the esophagus and the stomach; they consist in waves of contraction that progress along the intestine. There are two types of peristalsis: (*a*) slow waves, described by Bayliss and Starling,[2] which travel at the rate of 2 cm. per min. for a short distance; (*b*) fast waves, which progress at speeds up to 25 cm. per min. and travel a long way down the intestine, sometimes its whole length. The fast waves have been named *peristaltic rushes* by Alvarez.

Bayliss and Starling,[3] in their classic observations on peristalsis, came to the conclusion that slow waves are produced by a reflex caused by the distention of the intestine, in which Auerbach's plexus plays a part. They consist in contraction of the circular muscle above the distention, with relaxation in the immediately distal part of the intestine (*myoenteric reflex*). Contraction and relaxation simultaneously progress as a wave for a short distance. Stimulation of the intestine (*e.g.*, by pinching) is followed by contraction above and relaxation below the stimulated point. This reflex takes place even after cutting the extrinsic nerves of the intestine, but is suppressed by the local application of cocaine or nicotine. Bayliss and Starling believed this was the typical motor activity of the intestine (*the law of the intestine*). Later Alvarez found that this type of activity is the only one observed after denervation, but in the normal organism the other types of activity are more commonly seen. Peristaltic waves push forward the intestinal content. It has been proved that the bolus

[1] Alvarez, W. C., *Am. J. Physiol.*, **35**, 177, 1914.
[2] Bayliss, W. M., and E. H. Starling, *J. Physiol.*, **24**, 99, 1899.
[3] Bayliss, W. M., and E. H. Starling, *J. Physiol.*, **26**, 125, 1901.

progresses in a spiral fashion, probably because of the arrangement of the muscular fibers of the intestine.

Peristaltic rushes occur with greater frequency and travel a longer distance when the whole intestine is active than when it is in a quiescent state. The best way to stimulate a peristaltic rush is to drink a certain amount of fluid. Sometimes swallowing movements are enough to start a rush; at other times, the passage of chyme into the duodenum or a peristaltic wave from the stomach starts a peristaltic rush in the small intestine. This type of movement is apparently due to a reflex.

Pendular and other types of movement. Pendular movements are due to annular constrictions that travel up and down a short length of the bowel. These movements are repeated with a frequency of 10 to 20 per minute; they do not push the contents along the intestine, but serve to mix them with the digestive juices and to bring them into contact with the intestinal villi.

Beside the three principal types of movement already described (segmentation, peristalsis, and pendular movements) there are others, such as tonic waves, which progress slowly along the intestine, a contraction of systolic type that occurs only in special conditions, and antiperistaltic waves.

Coordination of intestinal movements. Little is known about the mechanisms which integrate and coordinate the movements of the small intestine in such a manner that one type of movement follows another in an orderly sequence, thus assuring the normal progress of digestion and absorption. Many years ago Cannon[1] observed that foodstuffs of different kinds advanced along the intestine at different rates, yet digestion and absorption were equally complete when the intestinal residue arrived at the final part of the ileum. Gregory[2] has studied this problem in dogs with a Thiry-Vella fistula in the jejunum, after the animals had recovered from the operation, so that they could be examined without anesthesia. After a prolonged period of fasting the intestinal loop was in a "basal" condition in which alternating periods of activity and quiescence were observed. After feeding, tone and activity were increased for several hours. This is a reflex response, mediated

[1] Cannon, W. B., *Am. G. Physiol.*, **12**, 387, 1904.
[2] Gregory, R. A., *J. Physiol.*, **111**, 119, 1950.

by the vagi, which arises in (*a*) cephalic receptors (sham feeding provokes it), and (*b*) the stomach, when it is distended by food. The composition of the intestinal content also conditions the tone and movements of the loops: bile salts, the products of protein digestion, hypertonic saline, and acids increase tone and motility. These effects are not suppressed by denervation or by blocking the ganglia with tetraethylammonium. Application of a local anesthetic suppresses the response, which seems, therefore, to be dependent on local nervous activity in the intestinal wall.

Polarization. Antiperistaltic waves are frequently observed in the duodenum and also in the last segment of the ileum; in other parts of the intestine they are seldom seen. Mall[1] and others have reported that if a sufficiently long segment of intestine is cut out and then reunited to the intestine with the ends inverted, this segment becomes a barrier that stops the progress of the intestinal contents in spite of the continuity and patency of the intestinal canal. The intestine is therefore "polarized" in the sense that fluids and solids are made to progress in one direction only. The cause of this polarity, according to Bayliss and Starling, resides in the "law of the intestine"—the fact that when a stimulus acts on the intestine there is contraction above the point stimulated and relaxation below. Alvarez maintains that the metabolic gradient is the underlying factor of intestinal polarity.

The action of drugs.[2] Drugs modify intestinal tonus and peristalsis. Atropine, banthine,[3] and morphine diminish, and prostigmine and pitressin increase, the strength of contraction in the jejuno-ileum of man. The first three drugs retard the passage of the contents along the gut; the last two accelerate it.

By means of radiologic studies in dogs it has been demonstrated that ergotamine and prostigmine accelerate the passage along the intestine, while atropine and ephedrine retard it.[4]

The influence of innervation. Whatever may be the cause of intestinal polarity, it is independent of the extrinsic innervation. Intestinal movements are not significantly modified by the complete denervation of the small intestine. Rhythmic segmentation is myogenous, and slow peristalsis is apparently due to reflexes in the Auerbach plexus.

Auerbach's plexus lies between the circular and longitudinal muscle layers. It has many ganglia and connections with Meissner's plexus, which is situated in the submucosa. The vagus nerve ends on the ganglion cells of Auerbach's plexus. Stimulation of the vagus produces, after a short period of inhibition, an increase in tonus and peristalsis. Section of the vagus decreases intestinal tonus but has no influence on peristalsis; therefore its action is not continuous on this last type of movement.

The majority of sympathetic fibers going to the intestine travel in the minor splanchnic nerves to the celiac and superior mesenteric ganglia, where the postganglionic fibers to the intestine commence. Stimulation of the sympathetic, or injection of adrenaline, inhibits peristalsis and diminishes intestinal tonus. Section of the splanchnic nerves is followed by an increase in peristalsis; therefore the sympathetic has a continuous inhibitory action on this type of motility. There is some proof that ileus paralyticus (intestinal paralysis) occurring in peritonitis is due to stimulation of the splanchnics.

The Movements of the Duodenum

The movements of the duodenum deserve special consideration. The duodenal cap is passively distended by the chyme that passes through the pylorus. The duodenal cap then contracts, and its contents are passed on to the second and third parts of the duodenum. These parts have peristaltic movements that send the chyme forward, but there are also antiperistaltic movements that send it back into the duodenal cap and even into the stomach. Once the contents have reached the fourth part of the duodenum, they do not go back.

The Emptying Time of the Small Intestine

The chyme reaches the last part of the small intestine in about three and a half hours after the commencement of gastric evacuation. It stays there for a certain time, and rhythmic segmentation of the bowel helps the absorption

[1] Mall, F. P., *Johns Hopkins Hosp. Rep.*, **1**, 93, 1896.

[2] Montero, E., F. Huidobro, and F. Cuevas, *Rev. de med. aliment.*, **6**, 14, 1943–1944; Ingelfinger, F. J., *New England J. Med.*, **229**, 114, 1943; Miller, I. G., *Gastroenterology*, **3**, 141, 1944.

[3] Chapman, W. P., *et al.*, *J. Clin. Investigation*, **29**, 803, 1950.

[4] Van Liere, E. J., D. W. Northup, and J. C. Stickney, *Am. J. Physiol.*, **141**, 462, 1944.

of foodstuffs. Peristalsis is not very active in this last part of the ileum, but it is increased by the ingestion of food (gastroileal reflex). Thus the contents pass on to the colon.

Where the ileum ends in the cecum, for a length of about 2 cm. the muscular layer thickens and forms a sphincter. This ileocolic sphincter relaxes before a peristaltic wave and then contracts, thus avoiding regurgitation of the contents of the cecum into the ileum. Usually the sphincter resists a pressure of 40 to 50 cm. H_2O, but in some cases it is overcome by pressures lower than these (incontinence).

The sphincter is not controlled by the vagus. Stimulation of the splanchnic nerve provokes strong contraction of the sphincter.

The ileocecal sphincter has a double function; it prevents the passage of the contents of the cecum into the ileum, and it prevents a too-rapid emptying of the ileum into the cecum, thus giving time for complete digestion and absorption.

BIBLIOGRAPHY

ALVAREZ, W. C., "An Introduction to Gastroenterology," 4th ed., Hoeber, New York, 1948.

BARCLAY, A. E., "The Digestive Tract. A Radiological Study of Its Anatomy, Physiology and Pathology," Cambridge, New York, 1933.

CANNON, W. B., "The Mechanical Factors of Digestion," E. Arnold & Co., London, 1911.

DÉLÉZENNE, C., La sécrétion interne du pancréas, in G. H. Roger and L. Binet, "Traité de physiologie normale et pathologique," Vol. 2, Masson et Cie, Paris, 1931, p. 294.

FLOREY, H. W., R. D. WRIGHT, and M. A. JENNINGS, The Secretions of the Intestine, *Physiol. Rev.*, **21**, 36, 1941.

HORRALL, O. H., The Toxicity of Bile, *Physiol. Rev.*, **11**, 122, 1931.

IVY, A. C., The Physiology of the Gall Bladder, *Physiol. Rev.*, **14**, 1, 1934.

LA BARRE, J., "La sécrétine," Masson et Cie, Paris, 1936.

LAGERLÖF, H. O., "Pancreatic Function and Pancreatic Disease Studied by Means of Secretion," Macmillan, New York, 1942.

MANN, F. C., The Functions of the Gall Bladder, *Physiol. Rev.*, **4**, 251, 1924.

NEWMAN, C., The Physiology of the Gall Bladder, *Lancet*, **1**, 785, 1924.

NORTHROP, J. H., "Crystalline Enzymes," Columbia University Press, New York, 1939.

ROGER, G. H., Physiologie du foie, in G. H. Roger and L. Binet, "Traité de physiologie normale et pathologique," 2d ed., vol. 3. Masson et Cie, Paris, 1939, p. 1.

SCHMIDT, C. L A., The Extrahepatic Function of Bile, *Physiol. Rev.*, **7**, 129, 1927.

STILL, E. M., Secretion, *Physiol. Rev.*, **11**, 328, 1931.

THOMAS, J. E., "The External Secretion of the Pancreas," Charles C Thomas, Springfield, Ill., 1950.

CHAPTER 39

The Large Intestine

THE LARGE INTESTINE[1] consists of the cecum and the appendix, the colon, the rectum, and the anal canal. Its principal function is to form, transport, and evacuate the feces. The shape and size of the colon differ considerably from one species to another. Carnivora have a short colon and a small cecum; the latter in some cases is missing. Herbivora have a long colon and a large cecum full of foodstuffs. This difference is due to the fact that in herbivora the bacteria of the large intestine play an important part in digestion; they dissolve the cellulose covering of the food and thus set free the nutritive contents. In carnivora digestion and absorption are almost complete by the time the contents arrive at the cecum. The large intestine can absorb only water, glucose, and salts. The large intestine of man resembles that of carnivora in length, but it has the haustral segmentation seen in herbivora.

Compared with other parts of the digestive tract the large intestine differs by its capacity, its distensibility, and the fact that its contents are held back for a considerable time.

Absorption, secretion, and excretion. One of the important functions of the large intestine is the absorption of water. The contents of the cecum and the ascending colon are almost liquid. The greater part of the remaining water content and some soluble substances are absorbed there, but the capacity to absorb fluids and crystalline substances exists all along the large intestine, down to the sigmoid and rectum. This property is used in therapeutics to give by rectum water, minerals (salt, calcium, iron), carbohydrates (glucose), and drugs (atropine, mercurial diuretics, salts of mercury, etc.).

The large intestine secretes a thick alkaline fluid (pH 8.4), rich in mucus. A more fluid secretion is obtained by mechanical irritation of the bowel, stimulation of the pelvic nerve, and injection of pilocarpine.

Stimulation of the sympathetic and injection of atropine inhibit this secretion. The juice secreted by the large intestine has no enzymes and its function appears to be one of lubrication, thus facilitating the progress of the feces and protecting the mucosa.

The large intestine has an excretory function; thus heavy metals (Ca, Mg, Hg) and certain drugs used in therapeutics (*e.g.*, arsenicals) are excreted in great part through the large intestine.

Innervation of the large intestine. Two parts must be distinguished in the large intestine, with respect both to its innervation and to its movements: (*a*) a proximal part (the cecum, the ascending colon, and part of the transverse colon); (*b*) a distal part (left half of the transverse colon, the descending colon, and the ileopelvic colon).

The proximal colon is innervated by sympathetic and parasympathetic fibers from the superior mesenteric plexus; the former are inhibitory and the latter motor. The sympathetic fibers for the distal colon arise in the upper lumbar roots and travel through the intermesenteric and inferior mesenteric plexuses and their pelvic splanchnic branches. The parasympathetic fibers originate in the second, third, and fourth sacral nerves and reach the colon through the hypogastric plexus.

According to the findings of Ivy and his coworkers, the nerve fibers that go to the colon through the pelvic nerve are cholinergic, and those which travel in the hypogastric nerve are adrenergic.

The movements of the large intestine. The large intestine is loose in the abdomen so that its position varies with posture, with respiratory

[1] HARDY, T. L., *Lancet*, 1, 519, 1945.

movement, and with the degree of its distention. The more movable parts are the transverse and pelvic colon. The cecum, the ascending colon, and the hepatic flexure are less movable. The splenic flexure, the descending colon, and the rectum are relatively fixed structures.

The time taken by the contents of the intestinal tract to reach the ileocolic sphincter is extremely variable (30 min. to 5 hr.). Approximately 3½ hr. after gastric evacuation has begun, or even before if a second meal is taken (gastroileal reflex), the contents of the small intestine commence to pass into the cecum. The cecum and the ascending colon are passively filled with a fluid mass, which pours through the ileocolic sphincter in small jets. In 24 hr. approximately 450 cc. of water enters the cecum; the feces of a normal subject contain about 80 cc. of water and 35 gm. of solids; therefore the proximal part of the colon absorbs around 370 cc. of water.

In some animals (cat, rabbit) peristaltic movements, which carry the contents up into the first half of the transverse colon, are seen, followed by antiperistaltic movements which send them back into the cecum. In man the movements of the colon[1] are less frequent, and therefore difficult to observe by x-rays. There are two types of movement in the large intestine, propulsive and nonpropulsive movements, beside changes in tonus. Contraction of one segment makes the contents go forward only when the adjacent distal segment is relaxed; the length traveled by the propulsive wave depends on the coordination of adjacent segments. Extracts of the posterior lobe of the hypophysis (vasopressin, pitressin), prostigmine, and ergotamine increase the propulsive motility of the colon. Morphine increases the tonus and the nonpropulsive motility, but diminishes the propulsive activity. Atropine and banthine inhibit the motility of the colon. Usually the movements of the large intestine are few, and whatever their type, the feces make little or no progress. These movements churn the contents and help the absorption of water. Nevertheless, a few times a day, after a meal (gastrocolic reflex) or in the course of defecation, a strong contraction begins in the transverse colon and travels in a wavelike manner toward and down the descending colon, propelling into the pelvic colon the whole contents of the distal part of the large intestine.

[1] Atkinson, A. J., H. F. Adler, and A. C. Ivy, *J. A. M. A.*, **121**, 646, 1943.

A meal containing a substance opaque to x-rays, *e.g.*, barium, can be followed down the digestive tract, and the time taken to reach the different parts noted.[1] Usually 3½ hr. after the ingestion of the opaque meal the cecum begins to fill—a process which is completed by the seventh hour. The cecum and the ascending colon are filled, but the mass does not go farther than the hepatic flexure. Between the seventh and the eighth hours the transverse colon is filled as far as the splenic flexure. This flexure apparently presents an obstacle to the advance of the intestinal contents. The mass that fills the cecum and the ascending and transverse colon remains in these segments up to the twentieth hour after ingestion. During this time nonpropulsive movements are observed which churn the contents and facilitate the absorption of water. Around the twentieth hour a strong contraction of the colon begins. The transverse colon, which hangs loose in the abdomen, straightens out, the splenic flexure is rounded, and the contents are moved into the descending colon (mass peristalsis). All these movements last 1 to 4 sec. The colon from below the hepatic flexure (cecal sphincter?) up to the splenic flexure is thus emptied. Sometimes there is no mass peristalsis and the transverse colon empties itself in several stages, at more or less long intervals.

The opaque mass then fills the descending colon and sometimes part of the sigmoid. Around the twenty-fourth hour after a meal there is mass peristalsis in the descending and pelvic colon, and the rectum is filled. Defecation follows soon after (Fig. 179).

DEFECATION

Usually the feces are stored in the pelvic colon until defecation takes place. At the junction of the pelvic colon and the rectum there seems to be an obstacle to the progress of the feces; according to some observers there is a definite sphincter at this level (pelvirectal sphincter). Only when this obstacle is overcome by a movement of mass peristalsis and the feces distend the rectum (a pressure of 30 to 40 mm. Hg is enough) is a perineal sensation felt, which awakens the desire to defecate. The weight of the feces in the sigmoid does not provoke this

[1] Becker, R., and A. Oppenheimer, "Normale und pathologische Funktionen der Verdauungsorgane im Röntgenbilden," Thieme, Leipzig, 1931.

desire, but gives a sensation of fullness localized above the pubis.

Defecation is a reflex. Distention or stimulation of the rectum by movements of the feces within it or in the anal canal, or by other means, provokes reflexly a strong wave of mass peristalsis, contraction of the rectum, and relaxation of the internal and external anal sphincters, thus emptying the large intestine from the transverse colon down. This emptying is helped by the simultaneous contraction of the diaphragm and of the abdominal muscles, which increases abdominal pressure, and that of the levatores ani and other perineal muscles, which sustain the perineum and help to evacuate the last remaining portions of the rectal contents.

The afferent paths of the reflex travel along the pudendal and pelvic nerves. The efferent paths belong to the sympathetic and parasympathetic. The sympathetic fibers arise in the lumbar segments of the spinal cord and end in the inferior mesenteric and hypogastric ganglia, whence the postganglionic fibers go to the bowel in the lumbar colonic and hypogastric or presacral nerve. The impulses traveling in these fibers have a continuous inhibitory effect on the walls of the distal colon, and a motor effect on the internal anal sphincter. The activity of the distal colon increases when these fibers are cut; when they are stimulated, the colon relaxes. Stimulation also increases the tonus of the internal anal sphincter. The parasympathetic fibers arise in the sacral segments of the cord and form the pelvic nerve. They have a continuous motor effect. Section of the pelvic nerve relaxes, and stimulation increases, the motility of the distal colon. Stimulation of the pelvic nerve relaxes the internal anal sphincter. The external anal sphincter is a striated muscle under voluntary control; it is innervated by fibers of the pudendal nerve that originate in the sacral segments of the cord.

The spinal center of defecation is situated in the lumbosacral segments of the cord, principally in the second, third, and fourth sacral segments. Section of the cord above this level does not disturb defecation once the effects of spinal shock have passed (see Chap. 81). Destruction of the sacral segments is followed by fecal incontinence; nevertheless in some cases after a time a certain degree of automatic control of evacuation is recovered. A medullary center, situated in the floor of the fourth ventricle, near the respiratory and the vomiting centers, has

FIG. 179. Diagrams of the movements of the colon: *a*, roentgenographic picture 5 to 6 hr. after ingestion of a barium meal; *b*. 7 to 8 hr.; *c*, 12 to 20 hr.; *d*, around 20 hr.; *e*, 20 to 24 hr.; *f*, 24 to 26 hr. after the barium meal. The dotted line shows the direction of the movement and the amount of previous filling. (*Becker, R., and A. Oppenheimer, "Normale und pathologische Funktionen der Verdauungsorgane im Röntgenbilden," Thieme, Leipzig, 1931.*)

been described. The spinal centers that control the tonus of the sphincters can be inhibited by peripheral stimulation or by intense emotions (fear, etc.).

From the second year of life in human beings the reflex mechanism of defecation is under voluntary control, at least in the ordinary conditions of life.

The Feces

In man and carnivora the ingested foodstuffs are almost completely absorbed before they reach the cecum. The semiliquid mass that passes through the ileocolic sphincter consists mainly of the residue of the digestive secretions (gastric, pancreatic, and intestinal juices; mucus; and bile), epithelial cells, and leukocytes. A great part of the water in the contents that pass into the cecum is absorbed in the large intestine, where bacteria and certain solid substances are added. A semisolid paste is formed, called the feces or fecal matter.

Fecal matter is therefore more a product of the activity of the digestive tract than a residue of foodstuffs. The following facts are proof of this assertion: (*a*) there are no soluble carbohydrates or proteins in the feces; (*b*) after a meat meal there are no muscle fibers in the feces; (*c*) the nature of the diet, if it does not contain cellulose, has no influence on the chemical composition of the feces; (*d*) in complete fasting fecal matter is produced and its chemical composition is the same as in normal feeding; (*e*) if a loop of the bowel is tied between two ligatures, after a time it is filled with a soft pastelike substance similar to fecal matter.

Fecal matter is made up of water (65 to 75 per cent); nitrogen (3 to 5 gm. per cent), mostly bacterial and endogenous; lipids (lecithin, coprosterol, fatty acids, and a verys mall amount of neutral fats), also mostly endogenous; a large number of bacteria; some minerals (Ca, P, Mg, Fe); and the products of disintegration of the bile pigments (stercobilinogen and stercobilin).

Cellulose ingested with the food is not absorbed in the small intestine; in man the bacteria of the large intestine do not have much effect on it. Nevertheless it plays an important part in digestion, because by increasing the bulk of the feces it stimulates peristalsis, accelerates the passage through the digestive tract, and helps intestinal evacuation.

THE MICROBIAL FLORA OF THE DIGESTIVE TRACT[1]

Many species of microorganisms inhabit the digestive tract. In the mouth there are numerous microbes: enterococci, staphylococci, spirochetes, spirils, vibriones, amebas, actinomyces, etc. The stomach and the proximal part of the small intestine are sterile or contain few microorganisms (mostly those of lactic fermentation); this is due to the antiseptic effect of the strong acidity of the gastric juice. The last part of the small intestine has some bacteria, among which *Escherichia coli* predominates; many of these come from the large intestine. Anaerobic bacteria do not grow in the ileum; therefore products of putrefaction are not found in this part of the intestine.

The number of microorganisms in the large intestine is enormous. Cohendy[2] has studied quantitatively the microbial flora of human feces and has found that 1 mg. of feces contains 143,870,000 microorganisms; according to him, bacteria make up about two-thirds of the dry feces.

The microbes in the digestive tract live on the protein and carbohydrate in the food, acting on them by two processes called putrefaction and fermentation. In putrefaction protein is attacked by the anaerobic bacteria and disintegrated into a series of substances that have the typical odor of putrefaction: amines (ethylamine, thyramine, histamine); volatile acids; phenols (indol and skatol) derived from tryptophane by deamination and decarboxylation; paracresol and phenol derived from phenylalanine and tyrosine by similar processes; substances containing sulfur (H_2S); and gases (H, CO_2, CH_4). In man and in carnivora, putrefaction is due almost exclusively to the anaerobic bacteria of the large intestine.

Carbohydrates are disintegrated by fermentation into a series of substances. Sugars are broken down to lactic acid in the stomach and the first part of the small intestine by the action of several germs. This process is of importance in suck-

[1] Lisbonne, M., Microbes et actions microbiennes dans le tube digestif, in G. H. Roger and L. Binet, "Traité de physiologie normale et pathologique," vol. 2, Masson et Cie, Paris, 1931, p. 445. Also papers by Scheunert and Schlieblich, Rietschel and Hummel, and Magnus-Alseben in Bethe's "Handbuch der normalen und pathologischen Physiologie," **3**, 967, 1001, and 1027, Springer, Berlin, 1927.

[2] Cohendy, M., *Compt. rend. Soc. de biol.*, **60**, 415, 1906.

lings. In the last part of the small intestine other microorganisms also attack sugars and break them down into acids, alcohols, ketones, and gases.

Microbes play an important part in the digestion of cellulose. None of the digestive enzymes acts as a cellulase for the hydrolysis of cellulose. Nevertheless, cellulose, the principal food of the herbivora, can be utilized thanks to the action of microorganisms. The digestion of cellulose in animals is a "monopoly of bacteria." In herbivora the architecture of the digestive tract (rumen, cecum, colon) allows a prolonged retention of the foodstuffs and thus facilitates the action of microorganisms. In the feces of man *Bacillus cellulosae dissolvens* has been found.[1] Although it has a potent activity on cellulose, very little cellulose is digested in the large intestine of man.

The microorganisms in the digestive tract sometimes play an important part in the development of pathologic conditions, but they are also prominent in certain physiologic functions; *e.g.*, their participation in the synthesis of some of the vitamins has been well demonstrated. Shortly after birth the intestinal flora commences to develop, and the blood-coagulation time becomes normal as a result of the synthesis of vitamin K by intestinal bacteria. In herbivora, intestinal bacteria also synthesize thiamine, riboflavin, pyridoxine, pantothenic acid, biotin, folic acid, and nicotinic acid. Bacterial activity is an important source of these accessory food factors, indispensable to animal life. Vitamins originated in the bowel are not always totally absorbed. Thus in rats fed a diet free from vitamin B_1 (thiamine), this substance is synthesized in the intestine but is not absorbed in adequate amounts. If the animals have access to their feces they eat them, and the symptoms of avitaminosis do not appear or are considerably delayed; this proves that vitamin B_1 is produced by the intestinal bacteria in amounts sufficient to keep the animals free from symptoms. But if coprophagia is prevented avitaminosis soon appears; therefore the thiamine produced is not immediately absorbed in sufficient quantities.

Recent work has shown that in adult human beings intestinal bacteria synthesize thiamine, biotin, folic acid, pyridoxine, pantothenic acid, para-aminobenzoic acid, vitamin B_{12}, nicotine amide, and riboflavin, and that this source of these vitamins is sufficient for the needs of the organism.[1]

[1] KHOUVINE, *Ann. Inst. Pasteur*, **37,** 712, 1923.

Intestinal bacteria not only synthesize some of the indispensable vitamins; they can also destroy them. Vitamin C is destroyed by many of the microorganisms in the bowel (*Escherichia*, *Aerobacter*, *Salmonella*, *Eberthella*, *Streptococcus encapsulatus*, and *Vibrio*) when there is anaerobiosis.[2] Nicotinic acid is also destroyed by intestinal bacteria in anaerobic conditions; but the destruction is not as rapid and complete as that of ascorbic acid (vitamin C).[3] There is also evidence of the destruction of vitamin B_1 (thiamine) in the human intestine. These facts explain why, in the treatment of some cases of vitamin deficiency, better results are obtained when the vitamin is given by injection than when it is given by mouth.

BIBLIOGRAPHY

BECKER, R., and A. OPPENHEIMER, "Normale and pathologische Funktionen der Verdauungsorgane im Röntgenbilden," Thieme, Leipzig, 1931.

GARRY, R. C., The Movements of the Large Intestine, *Physiol. Rev.*, **14**, 103, 1934.

GRACE, W. J., S. Wolf, and H. C. WOLFF, "The Human Colon," Hoeber, New York, 1951.

NEWTON, W. H., "Recent Advances in Physiology." Chapter 9, Defecation, J. A. Churchill, London, 1936.

OGILVIE, H., The Large Bowel and Its Functions, *Proc. Roy. Soc. Med.*, **44**, 200, 1951.

STARLING, E. H., Überblick über den gegenwärtigen Stand der Kenntnisse über die Bewegung und die Innervation des Verdauungskanal, *Ergebn. d. Physiol.*, **1**, 446, 1902.

TRENDELENBURG, P., Bewegung des Darmes, in Bethe's "Handbuch der normalen und pathologischen Physiologie," **3**, 452, Springer, Berlin, 1927.

WESTPHAL, K., Die Defäkation, in Bethe's "Handbuch der normalen und pathologischen Physiologie," **3**, 472, Springer, Berlin, 1927.

[1] NAJJAR, V. A., and L. E. HOLT, JR., *J. A. M. A.*, **123,** 683, 1943; NAJJAR, V. A., G. A. JOHNS, G. C. MEDAIRY, G. FLEISCHMANN, and L. E. HOLT, JR., *J. A. M. A.*, **126,** 357, 1944; NAJJAR, V. A., E. H. HOLT, G. A. JOHNS, G. C. MEDAIRY, and G. FLEISCHMANN, *Proc. Soc. Exper. Biol. & Med.*, **61,** 371, 1946; DENKO, C. W., *et al.*, *Arch. Biochem.*, **10,** 33, 1949; DYKE, W. J. C., *et al.*, *Lancet*, **1,** 486, 1950.

[2] YOUNG, R. M., and L. F. RETTGER, *J. Bact.*, **46,** 351 1943.

[3] BENESCH, R., *Lancet*, **1,** 718, 1945.

SECTION FIVE

Metabolism and Nutrition

CHAPTER 40

The Material and Energy Exchanges of the Body

Definitions. Metabolism consists of the exchange of matter and energy between the organism and the environment, and also the material and energy transformations within the organism. At one time the terms "metabolism" and "nutrition" were considered synonymous, but now "nutrition" signifies the relation existing between the ingestion of food and metabolism, *i.e.*, the sum of the processes concerned in the growth, maintenance, and repair of the living body as a whole or of its constituent organs (Lusk). "Metabolism" is restricted to the energy and material transformations within the body.

General description. Metabolic processes can be considered from two points of view:

1. The material cycle, *i.e.*, changes undergone by substances in the body in the different periods of life—during growth, the equilibrium of the adult, and the involution of senescence.
2. The energy cycle, *i.e.*, transformations undergone by the chemical energy in the foodstuffs, until it is finally eliminated as heat in the resting animal, or as heat and work when the muscles are active.[1]

There are two fundamental processes in the metabolism of matter: (*a*) anabolism, or assimilation; (*b*) catabolism, or the disintegration of substances. Assimilation is the formation of the specific substances of the organism from absorbed substances. Catabolism consists of the processes of disintegration of substances in the tissues into simpler ones; the end products ar usually excreted.

Life is maintained by a continuous transfor mation of matter and energy, which are taken i with the foodstuffs. Food fulfills the followin functions: (*a*) it provides raw material for th formation of the mass and the specific substance of the body needed for growth, maintenance and repair of the results of the incessant wear an tear of the tissues; (*b*) it provides energy, whic the organism transforms in the course of it activity;[1] (*c*) it provides substances that regulat metabolism (vitamins, etc.).

Food is first submitted to digestive processes which convert colloidal nondiffusible substance into crystalloid diffusible ones, and comple foreign substances into simple ones, easily ab sorbed, which can be used for the formation o the specific substances of the organism. Thes nutritive substances are then absorbed and cir culate in the organism. Some are used up im mediately by the tissues; others are stored, some times after undergoing certain processes tha change them into forms suitable for storage. I the course of metabolism, residual or waste prod ucts are formed which are eliminated by th organs of excretion (kidney, intestine, skin, lung)

An outstanding feature of metabolism is th tendency to maintain an equilibrium. Thus th basal oxygen consumption and the fasting blood sugar level do not vary much from day to day and if any change does occur, the initial rate o

[1] In man there is no significant output of energy in other forms such as occurs in some animals, *e.g.*, electricity in electric fishes, light in certain insects, etc.

[1] This energy is used up by basal metabolism, by th specific dynamic action of foodstuffs, in work per formed by the organism, and in maintaining the bod temperature.

level is rapidly reestablished. This stability of metabolic processes is due to a complex aggregate of nervous and humoral factors which regulate these processes.

Several methods are used in the study of the metabolism of a substance: (*a*) the substance is administered, and its excreted end products are examined by chemical methods; (*b*) the substance administered is especially prepared with an isotope of one of its constituent elements, easy to detect by some physical character (weight, radioactivity) so that it can be followed in its passage through organic fluids and tissues and in the secretions and excretions of the body; (*c*) an organ, isolated from the rest of the body, can be maintained alive and functioning by perfusing it with whole blood or a nutrient fluid, and its metabolism can thus be studied; (*d*) respiratory and other metabolic processes of surviving slices of tissue, or of ground tissue, can be studied in Warburg's and other similar apparatus.

The use of isotopes. Remarkable results are being obtained in the study of the metabolism of animals, plants, and bacteria by the continuously increasing use of isotopes.[1] Isotopes are atoms of chemical elements which have the same number of planetary electrons as the "normal" atom but more or less neutrons in the nucleus. They have the same chemical properties, because these are dependent on the number of electrons, which is the same in all the atoms. The physical properties, on the other hand, are different because the atomic mass of each isotope differs from that of the "normal" atom because of the different number of neutrons in the nucleus. There are natural and artificial isotopes. Some are stable; others are radioactive. Deuterium or "heavy" hydrogen is a stable isotope. Its atomic weight is 2 (twice that of "normal" hydrogen) because its nucleus has one proton and one neutron, instead of only one proton.

Radioactive isotopes can be prepared by bombarding an atom in a cyclotron, or by bombarding a "normal" atom with neutrons in an atomic pile so as to eliminate one or more neutrons from its nucleus. The radiation these isotopes emit can be β or γ rays, or positrons. By measuring the intensity of radiation in a Geiger-Müller counter or some other electrometer, it is possible to detect and to estimate the concentration of the isotope. The stability of a radioactive isotope is measured by the time taken for its radiation to lose half its initial activity; this is known as the "half life." Thus P^{32} has a half life of 14.3 days, and K^{42} one of 12.4 hr. Isotopes do not lose their physical characteristics by taking part in chemical reactions, therefore they can be followed in the different products resulting from the chemical processes of the organism in which the substances they form part of are active. The most frequently used isotopes in biochemical studies are C^{14}, I^{131}, Fe^{59}, N^{15}, Na^{24}, K^{42}, P^{32}, Cl^{38}, Ca^{45}, and S^{35}. They can be obtained through the United States Commission on Atomic Energy.

The use of isotopes has brought about revolutionary changes in the knowledge of the photosynthesis of chlorophyll, and of the metabolism of proteins, fatty acids, carbohydrates and minerals in animals, plants, and bacteria. The knowledge of the processes of intermediate metabolism is being rapidly developed by the use of isotopes.

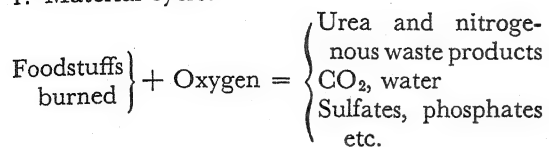
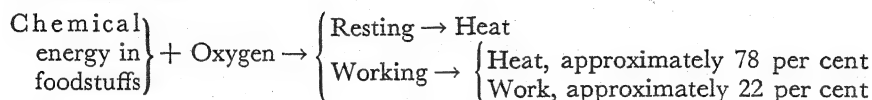
[1] BUCHANAN, J. M., and A. B. HASTINGS, *Physiol. Rev.*, **26**, 121, 1946; SACKS, J., *Chem. Rev.*, **42**, 411, 1948; HAHN, P. F., R. D. EVANS, and A. S. KESTON in Glasser's "Medical Physics," Year Book Publishers, Inc., Chicago, 1950, p. 641; KELSEY, F. E., *J. A. M. A.*, **146**, 1131, 1951.

Material and energy balanes. The study of general metabolism consists in establishing the balance of all the substances or all the energy that enters into the organism and is eliminated from it. On the other hand, the special metabolisms of one particular substance (*e.g.*, calcium, iodine, water) can be made the object of a particular study. There is equilibrium in metabolism (*e.g.*, calcium equilibrium) when the output of the substance is equal to the intake. A positive balance signifies that there is a gain in substance or energy because the intake is greater than the output; when there is a negative balance, there is a loss of substance or energy because the output exceeds the intake.

Foodstuffs have a twofold function. Some can be oxidized by the organism, thus releasing chemical energy, which is finally converted into heat and work; others are the raw material for the building and repairing of tissues and organic fluids. Organic foods, such as carbohydrates, proteins, and fats, serve to build up the specific substances of the body and to provide energy; inorganic foods, such as water and salts, do not yield energy but are necessary for the formation and upkeep of the body fluids and tissues.

The material balance of the organism is established by measuring the intake of foodstuffs and oxygen on the one hand and on the other the excreta eliminated in the urine, in the feces, and through the skin and lungs.

ESTIMATION OF THE MATERIAL AND ENERGY BALANCE

Methods. The composition of foodstuffs is determined by chemical analysis. When experiments must be made with animals fed on a standard diet, a large quantity is prepared and a sample is analyzed; the food is then stored, in the refrigerator if necessary, and used as it is needed. Foodstuffs have been analyzed and the results published in tables. For practical purposes health officers and clinicians can establish approximately the composition of any diet by reference to these tables. Protein is measured by estimating the total nitrogen content by the Kjeldahl method and multiplying the result by 6.25 (proteins have on an average 16 per cent N, therefore 1 gm. N = 6.25 gm. proteins). Fats are estimated by the weight of the ether extract. Carbohydrates are measured by submitting the foodstuffs to hydrolysis with dilute acid, and then measuring the amount of reducing substances; the result is usually given in terms of glucose. Water is measured by drying the food at 100 to 110°C. (or by distilling it); the difference in the weight before and after drying gives the water content. Ash content is determined by weighing after calcination. Methods for the analysis of foodstuffs are being continuously improved.

Nitrogen is excreted in the urine (90 to 98 per cent), in the feces (2 to 10 per cent),[1] and through the skin (insignificant amounts). Carbon is eliminated by the lung (CO_2) and in the urine and feces; oxygen and hydrogen by the lung, in the urine, in the feces, and through the skin; and salts in the urine, in the feces, and through the skin. All the energy at the disposal of animal organisms comes from chemical energy in foodstuffs. This energy is released in chemical reactions which lead to oxidation of C and H, with the formation of CO_2 and H_2O, which are eliminated from the body. In a burning candle, oxidation takes place directly and energy (heat, light) is immediately set free. In the cells, oxidation is performed in stages; hydrogen and electrons are transferred from one substrate to another by a series of enzymatic reactions in which the heat of combustion is gradually released (see Chap. 34, Cell Respiration).

Bioenergetics is based on a material and an energy cycle:

1. Material cycle:

Foodstuffs burned } + Oxygen = { Urea and nitrogenous waste products; CO_2, water; Sulfates, phosphates etc.

2. Energy cycle:

Chemical energy in foodstuffs } + Oxygen → { Resting → Heat; Working → { Heat, approximately 78 per cent; Work, approximately 22 per cent

The energy balance is established by measuring the energy entering the organism (caloric value of food ingested) and the output of energy (heat and work). Energy is measured in calories. A calorie is the amount of heat needed to raise the temperature of 1 gm. (small calorie, cal.) or 1 kg. (large Calorie or kilogram-calorie, kg.-cal.) of water from 14.5 to 15.5°C. Work is easily converted into calories; 427 kg.-m. is the equivalent of 1 kg.-cal. (1 kg.-m. = 0.0024 kg.-cal.). One kilogram-calorie (15°C.) is equivalent to 4.1855 joules (1 joule = 0.23892 kg.-cal.).

Table 33. Direct and Indirect Calorimetry in Human Subjects

Condition	*Duration of experiment, days*	*Average calories in 24 hr.*	
		Direct calorimetry	*Indirect calorimetry*
Resting..........	41	2246	2246
Working..........	66	4682	4676

Source: Atwater, W. O., *Ergebn. d. Physiol.*, **3**, 497, 1904.

There are three principal calorimetric methods: (*a*) direct calorimetry, in which heat eliminated is measured in calorimeters; (*b*) indirect respiratory calorimetry; (*c*) indirect calorimetry by establishing the energy balance. Two of these methods can be combined; the more complete respiratory calorimeters combine all three. Direct and indirect calorimetry give the sam-

[1] Approximately 1 gm. of N in the feces comes from the organism (digestive juices, etc.); the rest is due to undigested food and microorganisms.

results in both the resting and the working subject (Table 33).

The caloric value of foodstuffs is measured by burning them with oxygen in Berthelot's or Mahler's calorimetric bomb submerged in water. The rise in temperature of the water, multiplied by the amount of water, gives the calories; *e.g.*, if there is an increase of 6°C. in the 12 liters of water in which the bomb is submerged, the food burned has set free 72 kg.-cal. The caloric value of a food can also be measured by determining the amount of oxygen needed for the complete combustion of 1 gm. of the food, and converting this into calories.[1]

Direct Calorimetry

The first calorimetric determinations were made by Lavoisier and Laplace. They placed a guinea pig in a closed box surrounded by ice. The heat output of the animal was measured by the amount of water resulting from the ice melted. This heat corresponded to that set free by the combustion of an amount of carbon equivalent to the CO_2 eliminated by the animal. They were thus able to demonstrate for the first time that animal heat has its origin in combustions taking place within the body. Lavoisier and Laplace's calorimeter has only a limited use, because it measures the heat output only at an environmental temperature of 0°C.

[1] Benedict, F. G., and F. L. Fox, *J. Biol. Chem.*, **66**, 783, 1928.

A physiological calorimeter must fulfill the following conditions: (*a*) it should permit exact heat measurements at different temperatures; (*b*) there should be no accumulation of CO_2, nor decrease of oxygen; (*c*) it should be possible to study the influence of food and exercise. Direct calorimetry can be performed by means of (*a*) adiabatic calorimeters; (*b*) gradient calorimeters.

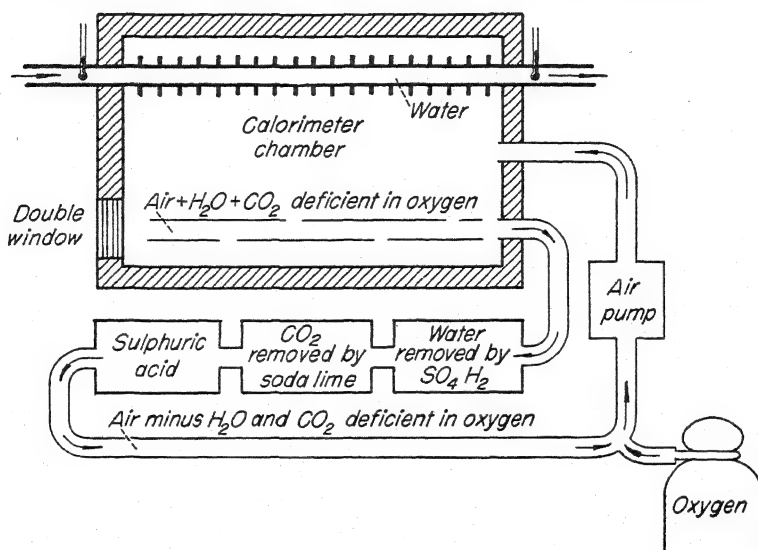



Fig. 180. Diagram of Atwater and Benedict's calorimeter.

An adiabatic calorimeter consists of a closed chamber with walls that do not allow the passage of heat to and from the environment (Fig. 180). Heat radiated by the subject (about three-fourths of the total heat production) is taken up by water circulating in a system of metal pipes within the calorimetric chamber. The increase in the temperature of the water as it passes through the pipes and the total amount of water are measured, thus obtaining the calories radiated by the subject. Water evaporated by the subject is collected in sulfuric acid and weighed; by multiplying the grams of water evaporated by 0.586, the caloric value is obtained. If the temperature of the subject varies in the course of the experiment, the degrees of temperature variation are multiplied by the body weight of the subject and by 0.83 (the specific heat of 1 kg. human body substance, with respect to the specific heat of water, which is 1). For example, if there has been a fall or a rise of temperature of 1° and the subject weighs 70 kg., 58.1 cal. should be subtracted from or added to the total heat loss. Direct calorimetry

is the standard method for measuring the energy converted in the body, but it is costly and complicated, requiring special equipment and trained personnel;[1] for this reason it is not frequently used.

through the lungs and in the urine and feces. Thus the material balance is established, and the energy converted by the subject into heat (when resting) or into heat and work (when working) can be calculated.

Table 34. Intake and Output of Nitrogen, Carbon, and Hydrogen in Man

Condition	Duration of experiment, days	Average daily intake in food, gm.		Average daily output, gm.					Gain or loss
				Feces	Urine	Water	Respiration	Total	
Resting	42	Nitrogen	18.5	1.1	17.8	....	.....	18.9	−0.4*
		Carbon	253.3	9.9	12.4	....	217.9	240.2	+13.1†
		Hydrogen	39.2	1.4	3.5	41.8	.....	46.7	−7.5
Fasting 1 to 2 days	5	Nitrogen	0	0	13.1	....	.....	13.1	−13.1‡
		Carbon	0	0	9.9	....	184.4	194.3	−194.3§
		Hydrogen	0	0	2.7	98.3	.....	101.0	−101.0

Source: ATWATER and BENEDICT, U.S. Department of Agriculture, Bull. No. 136, p. 120, 1903.
* Protein lost by the body: 2.5 gm. (calculated rrom N lost).
† Fat gained by the body: 19.2 gm. (calculated from C retained).
‡ Protein lost by the body: 82.0 gm. (calculated from N lost).
§ Fat lost by the body: 198.3 gm. (calculated from C lost).

Gradient calorimetry consists in measuring thermal output from a source of heat by enclosing it in a shell completely lined with a uniform heat-flow metering layer, while heat lost by ventilation and respiration is measured in heat-exchange meters based on the same principle. Calories conveyed to the heat-measuring gradient layers by radiation, conduction, convection, or evaporation and condensation are measured quantitatively by means of thermojunctions of copper and constantan. The difference in temperature of the different layers, their thickness, and their thermal conductivity are measured and expressed as heat flow in calories per second. Gradient calorimeters are simpler to handle than adiabatic calorimeters and quite as accurate, and the results are obtained very rapidly. They are therefore especially valuable for clinical and physiological work.[2]

INDIRECT CALORIMETRY BY DETERMINATION OF ENERGY BALANCE

Food ingested and oxygen consumption are measured; also the excretion of substances

[1] Among the most complete and exact calorimeters are those of Atwater, Rosa, and Benedict at the Carnegie Institution; of Lusk and Du Bois at the Russell Sage Institute of Pathology in New York; of Lefevre in Paris, etc.

[2] BENZINGER, T. H., and C. KITZINGER, "Direct Calorimetry by Means of the Gradient Principle," Naval Med. Res. Inst. Rep. No. 1, Bethesda, 1949.

In the experiment summarized in Table 34 with the resting subject, there is almost an equilibrium between the intake and output; there is a small loss of nitrogen (0.4 gm.) corresponding to 2.5 gm. protein; a certain amount of fat has been deposited, as is shown by the gain of 13.1 gm. carbon. In the experiment with the fasting subject, considerable loss of protein and fat was observed.

The following calculation is made: The amount of protein, carbohydrate, and fat in the feces is subtracted from that of the food ingested. Nitrogen is multiplied by 6.25 (the protein equivalent). If a subject in the course of 4 days takes in 15.1 gm. of N and eliminates 17.2 gm. (16.2 gm. in the urine, and 0.9 gm. in the feces), the difference, *i.e.*, 2.1 gm., has been lost by the body. The subject consumed 94.4 gm. of protein taken from the food and 13.2 gm. from the organism (a total of 107.5 gm. protein, equivalent to 17.2 gm. N). From the total carbon excreted, protein carbon ($N \times 3.2$) is eliminated; the rest corresponds to carbohydrate and fats. Carbohydrate is considered to be completely consumed; the carbon corresponding to carbohydrate ingested is subtracted from the total and the remainder corresponds to fat. Carbohydrate is obtained by multiplying C by 2.29 (carbohydrates contain 43.5 per cent C). The carbon content of fat is 76 per cent; therefore multiplying C corresponding to fat by 1.31, we have total fat. The caloric value of the protein, carbohydrate, and fat is then calculated.

Caloric value of foodstuffs. In 1842 Joule discovered the mechanical equivalent of heat, and shortly afterward Mayer (1843) and Helmholtz (1847) enunciated the law of conservation of energy, which states that the total energy of the universe is constant but that one form of energy (electrical, mechanical, chemical) can be converted into another. This law of conservation and conversion of energy is also valid in the energetics of animal organisms. This fact was demonstrated by Lavoisier, Pettenkofer and Voit, Rubner, and Benedict and Atwater, and it has been amply confirmed by many others.

Table 35. Calories Obtained from Foodstuffs

Foodstuff	Calories obtained	
	Experimental result	*Approximate practical value*
Fat, 1 gm.	9.3	9
Carbohydrate, 1 gm.	4.1	4
Protein, 1 gm.	4.1	4

Animals convert chemical energy bound in the foodstuffs into heat when resting and into heat and mechanical energy during work. Rubner measured the heat set free by the combustion of different foodstuffs and then determined the heat set free by the organism in the metabolism of those same foodstuffs. He found that fat and carbohydrate gave out the same amount of heat when burned in the calorimeter as in the organism, but that protein gave out less heat in the organism. This is due to the fact that urea and other metabolic waste products of protein metabolism are excreted by the body without being burned and the corresponding calories are lost to the organism, while there is complete combustion of protein in the calorimeter. Thus 1 gm. of protein sets free 5.3 kg.-cal. in the calorimeter and only 4.1 in the organism. On the other hand carbohydrate and fat set free 4.1 and 9.3 kg.-cal. respectively both in the calorimeter and in the organism. These are average values for the three kinds of foodstuffs. There are small differences between the different foodstuffs in each class; thus 1 gm. of starch sets free 4.2 kg.-cal., and 1 gm. of sugar 3.96 kg.-cal. The caloric value of animal proteins is greater than that of vegetable proteins. Taking into account the small loss of heat occurring in the course of digestion and the fact that not all the food ingested is absorbed, in practical dietetics the approximate values given in Table 35 are commonly applied.

At one time it was thought that 1 gm. of carbohydrate, 1 gm. of protein, and 0.45 gm. of fat, which give the same amount of heat, were isodynamic and could be interchanged one for the other. Further studies have considerably restricted the idea of isodynamia. When considering the metabolism of the different foodstuffs it will be seen that there is a minimum requirement of each of them, which cannot be replaced by an equivalent quantity of calories in another type of foodstuff.

Indirect Respiratory Calorimetry

Principle and methods. The amount of energy set free by the oxidation of foodstuffs in the organism can be calculated by measuring the oxygen consumption and the excretion of CO_2 during a certain time (Tables 37 and 38). There are three principal methods for determining respiratory exchanges: (*a*) confinement in a closed chamber; (*b*) the open-circuit method; (*c*) the closed-circuit method. In all these methods the whole animal can be placed within a respiratory chamber or the respiratory tract alone can be connected with a device for measuring its gaseous exchanges. In the latter case a mouthpiece, nasal catheter, or mask is used, taking good care to prevent leakage to or from the outside air. A system of valves which offer no resistance to breathing, but are perfectly competent, assures the circulation of air in one direction. The time over which the observation is made should be accurately measured, because the results are then expressed in terms of 1 hr. or 24 hr.

Confinement. Lavoisier used this method in his pioneer experiments in calorimetry. Its uses are limited to small animals and to observations of very short duration, because the decrease in oxygen and the accumulation of CO_2 soon disturb the normal functions of the animal.

Closed circuit. This method introduced by Regnault and Reiset in 1849 has since been frequently applied in different apparatus for metabolic determinations. The subject breathes air from and into a closed space; CO_2 and water are removed from the expired air by passing it through appropriate absorbents, and oxygen is added to the inspired air to replace what is absorbed by the organism. The method was ap-

plied in man in the Atwater-Rosa-Benedict respiratory calorimeter, for direct and indirect respiratory calorimetry. This apparatus permits the simultaneous measurement of heat production by both methods (Table 33). Air is made to circulate through the calorimetric chamber by means of a pump (Fig. 180). On leaving the chamber it passes through a flask containing sulfuric acid, which absorbs the water; then through a second flask containing soda lime, for the absorption of CO_2; then through a third flask with H_2SO_4, which absorbs water produced in the reaction between CO_2 and soda lime. The increase in weight of the first flask gives the subject's output of water; the increase in weight of the second and third flasks gives the CO_2 output. Oxygen absorbed by the subject is replaced from a tank; weight lost by this tank gives the oxygen consumption. The calorimetric chamber is large enough to allow the subject to live in it for several days; is furnished with a bed, a table, a telephone for outside communication, and a bicycle ergometer which accurately measures energy spent in work. This calorimeter has been used to give a precise demonstration that in man the energy output at rest (in the form of heat) or when working (heat and mechanical energy) is equivalent to the energy set free by the combustion of foodstuffs in the organism (Table 33). Direct and indirect calorimetry give equal results if performed for a sufficiently long period; during short periods there are sometimes certain differences. Thus during shivering and when the body temperature rises, the oxygen

FIG. 181. Diagram of Benedict-Roth recording spirometer.

consumption is greater than the output of heat. This is due to the retention of heat by the organism while the body temperature rises.

In medical practice the only method widely used is indirect respiratory calorimetry, because it is accurate and simple and does not require costly apparatus. A recording spirometer such as the Benedict-Roth model is used (Fig. 181). The subject breathes in and out of the spirometer, which is filled with oxygen. A set of valves directs the expired air into a receptacle with soda lime where CO_2 is absorbed. At each respiratory movement the bell of the spirometer rises and falls, tracing a continuously descending line as the subject takes up oxygen. The records (Fig. 182) are usually taken over a period of 6 min. and the result is calculated for 1 hr.

FIG. 182. A satisfactory record of oxygen consumption.

Knipping's apparatus measures oxygen consumption and CO_2 elimination. Expired air is breathed into a spirometer, and a pump then passes it into a flask containing KOH to absorb the CO_2, whence it is breathed in by the subject. The fall of the bell of the spirometer gives the oxygen consumption; the spirometer can be refilled with O_2 if a more prolonged observation is desired. The apparatus is then closed, and H_2SO_4 is poured into the flask containing KOH; CO_2 breathed out by the subject and absorbed by KOH is set free into the spirometer, the bell rises and the volume of CO_2 is thus measured.

Open circuit. In this method air or oxygen is breathed from the atmosphere, or from a tank; the expired air is measured and released into the atmosphere.

Haldane's apparatus (Fig. 183) for small animals (mice, rats, guinea pigs) gives excellent results.[1] The animal is placed in a small chamber, and air is made to circulate through the whole system by means of an aspirator or pump. Incoming air passes through soda lime and H_2SO_4, which absorb CO_2 and water. Outgoing air passes through sulfuric acid, where it is dried; then through soda lime, which absorbs CO_2; and finally through H_2SO_4, which absorbs water set free in the combination of CO_2 with soda lime. The flasks on the outgoing side are weighed at the beginning and end of the experiment. The increase in weight in the first two flasks containing H_2SO_4 gives the elimination of water; the

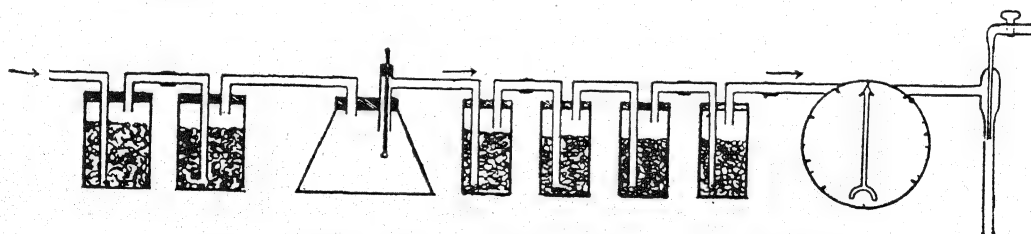



FIG. 183. Haldane-Pembrey respiration apparatus for small animals.

[1] HALDANE, J. S., *J. Physiol.*, 13, 419, 1892.

increase in weight in the other two flasks (soda lime and H_2SO_4) gives the CO_2 eliminated by the animal. Oxygen consumption is obtained indirectly by the following calculation: During the experiment only CO_2 and water are produced, and only oxygen is absorbed. The chamber with the animal is weighed at the beginning and end of the experiment. The difference is subtracted from the weight of CO_2 and water eliminated; the result corresponds to the weight of oxygen absorbed.

In man a Douglas bag (Fig. 184) strapped on the back can be used for collecting the expired air from subjects in many different physiologic conditions. A three-way stopcock (*B*) permits the expired air to pass into the bag or into the atmosphere. The subject breathes through a mouthpiece (1) into a set of perfectly competent valves (2) which direct the expired air into the rubber pipe going to the bag. The stopcock (3) is at first turned so that the expired air is blown out into the air; at a given moment it is turned so that the expired air goes into the bag during an accurately measured time. The bag is then closed by turning the stopcock and taken to the laboratory. The expired air in the bag is measured by passing it through a gas meter, and a sample is drawn into a Haldane buret.

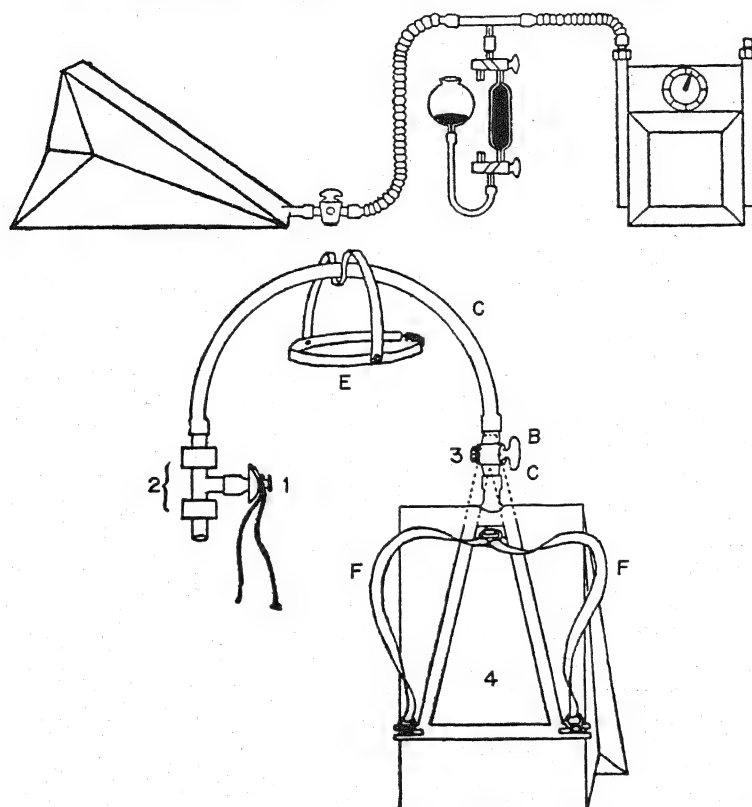



Fig. 184. Douglas bag for determining respiratory exchanges in man. Below: 1, mouthpiece; 2, inspiratory and expiratory valves; 3, stopcock for communicating the bag with the atmosphere or with the mouthpiece. Above: the bag is connected by a rubber tube with a gas meter; a sample of expired air is collected in the buret filled with mercury.

Instead of a Douglas bag, an accurately calibrated and balanced Tissot spirometer[1] can be used (Fig. 185). The subject breathes through a mask or mouthpiece and the expired air is directed by a set of valves into the spirometer, where it is measured. A sample is collected, and CO_2 and oxygen content are determined.

In the Haldane buret (Fig. 186) a small amount of expired air is accurately measured. CO_2 is absorbed by passing the air several times through a flask con-

[1] Boothby's, Bailey's, and other models can also be used.

taining KOH. The sample is again measured and the decrease in volume corresponds to the CO_2. Then the air is passed into a flask with alkaline pyrogallate, where the oxygen is absorbed. The sample is measured and the decrease in volume corresponds to the O_2. The composition of expired air is thus known, and

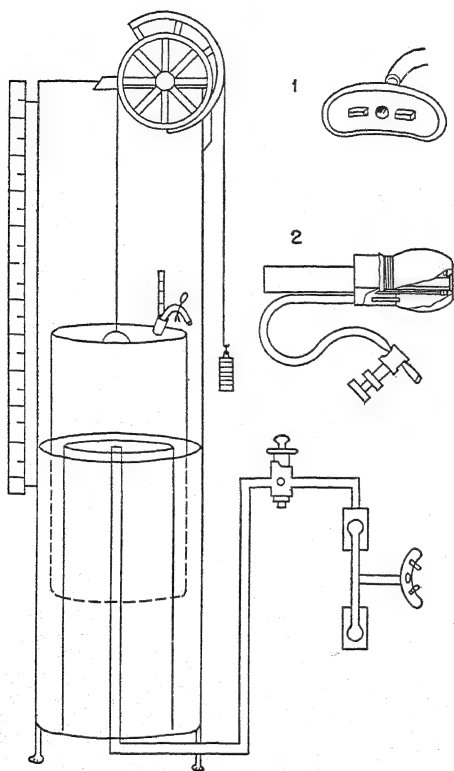



FIG. 185. Diagram of Tissot's spirometer. 1, mouthpiece; 2, nasal catheter. Expired air is sent through a valve into the bell of the spirometer, which is accurately calibrated and well balanced so that it does not offer any resistance to expiration, nor does it compress the gas collected.

as the composition of external air is constant, O_2 absorbed and CO_2 eliminated by the organism are easily calculated.

The Tissot-Boothby method, using the Haldane buret for the determination of respiratory metabolism, is very accurate but requires experience in gas analysis. The Benedict-Roth method is much simpler and is therefore more commonly used in medical practice. In both these methods care should be taken that the subject be physically and mentally at ease and his breathing regular. For this it is important that the mask or mouthpiece used should be comfortable and the valves and bell of the gasometer should not offer any resistance to breathing. The great drawback in these methods is the short time during which the sample is collected; as the results of a few minutes are expressed in terms of 1 hr., any transitory disturbance in respiratory metabolism or any error committed is multiplied several times. Far more reliable results are obtained when the observation lasts several hours or days, as when respiratory calorimeters are used. Working with two Tissot spirometers, it is possible to prolong the observation for a fairly long time, switching from one to the other and taking samples of expired air alternately from each gasometer.

Respiratory quotient.[1] The ratio of the volume of CO_2 eliminated to the volume of O_2 absorbed is known as the respiratory quotient (RQ). The volumes of the gases are considered

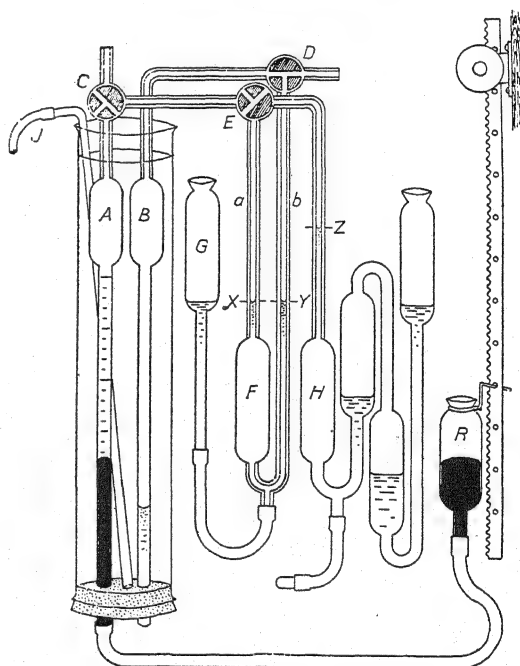



FIG. 186. Haldane's apparatus for gas analysis. *A*, buret for measuring the sample of air or gas to be analyzed; *F*, pipet containing KOH for absorption of CO_2; *H*, pipet containing potassium pyrogallate for absorption of oxygen; *R*, mercury reservoir for shifting the gas.

because according to Avogadro's law one molecule of O_2 has the same volume as one molecule of CO_2.

If foodstuffs are burned in oxygen *in vitro*, the RQ varies according to the foodstuff; therefore

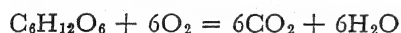
[1] RICHARDSON, H. B., *Physiol. Rev.*, **9**, 61, 1929.

the same variations must occur when they are burned in the organism.

The general formula of carbohydrates may be expressed as $C_n(H_2O)_n$; therefore in the course of combustion all the oxygen combines with C and forms CO_2. For example, in the combustion of glucose,

Table 36. Composition of Meat and Its Excreta

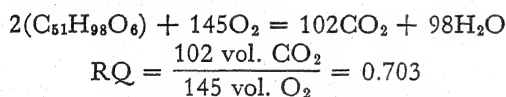
	C, *gm.*	*H*, *gm.*	*O*, *gm.*	*N*, *gm.*	*S*, *gm.*
Amount contained in 100 gm. meat	52.38	7.27	22.68	16.65	1.02
Residue eliminated in urine and feces	10.88	2.87	14.99	16.65	1.02
Balance used in combustion	41.50	4.40	7.69		

$$C_6H_{12}O_6 + 6O_2 = 6CO_2 + 6H_2O$$

The RQ will be equal to 1, because

$$RQ = \frac{6 \text{ vol. } CO_2}{6 \text{ vol. } O_2} = 1$$

In the combustion of fats O_2 is used up to oxidize not only carbon but also hydrogen. The RQ is therefore less than 1; for the fats in foodstuffs it is approximately 0.7. For example, in the combustion of tripalmitin,

$$2(C_{51}H_{98}O_6) + 145O_2 = 102CO_2 + 98H_2O$$

$$RQ = \frac{102 \text{ vol. } CO_2}{145 \text{ vol. } O_2} = 0.703$$

In the case of protein it is necessary to consider that some of the end products are eliminated in the urine and feces without being completely oxidized. C and H in the rest of the molecule are oxidized and eliminated as water and CO_2 (Table 36); for this purpose 138.18 gm. of O_2 is absorbed and 152.17 gm. of CO_2 is eliminated. As 1 gm. of O_2 = 0.6998 liters, and 1 gm. of CO_2 = 0.5094 liters, the following RQ results:

$$RQ = \frac{77.52 \text{ liters } CO_2}{96.7 \text{ liters } O_2} = 0.802$$

When carbohydrate is converted into fat, CO_2 is set free without the absorption of O_2; therefore the RQ is more than 1. If, on the contrary, fat is converted into carbohydrate, the consumption of O_2 for this purpose, without the corresponding elimination of CO_2, would cause the RQ to fall to approximately 0.3.

The RQ indicates which substances are being burned in the organism (Table 37). It must be remembered, however, that the figures given refer to each kind of foodstuff separately and that the body never burns exclusively one kind of substance, but the three principal kinds simultaneously; therefore the RQ is the result of many different chemical processes. Even when an excess of carbohydrate (*e.g.*, glucose) is given by mouth or in intravenous injection, the RQ rises only to 0.87; therefore other substances beside sugar are being burned.

When sugar is given to a patient with a severe diabetes, the RQ, which in these cases is 0.7, does not rise. It has been generally accepted that this means sugar cannot be burned in this condition, but this is not the only possible interpretation,[1] because an RQ of 0.7 may result from two processes that occur simultaneously and have opposite effects on the RQ; *e.g.*, sugar may be burned giving an RQ of 1, and at the same time carbohydrate may be converted into fat (there is as yet no final proof that this occurs), giving an RQ of 0.3.

Table 37. Gaseous Exchange, RQ, Heat Production, and Caloric Value of CO_2 and O_2 in the Metabolism of Common Foodstuffs

Substance	*O_2 needed to oxidize 1 gm., cc.*	*Produced in the oxidation of 1 gm.*		*RQ*	*Heat produced, cal.*	
		CO_2, cc.	*Heat, kg.-cal.*		*Per liter of O_2 consumed*	*Per liter of CO_2 consumed*
Starch	829.3	829.3	4.20	1.00	5.06	5.06
Saccharose	785.5	785.5	3.96	1.00	5.04	5.04
Glucose	746.2	746.2	3.74	1.00	5.01	5.01
Lactic acid	745.9	745.9	3.62	1.00	4.85	4.85
Animal fat	2013.2	1431.1	9.50	0.71	4.72	6.64
Protein	956.9	773.8	4.40	0.81	4.60	5.69
Acetone	1542.9	2157.2	7.43	0.75	4.82	6.42
Alcohol	1459.5	972.9	7.08	0.67	4.85	7.28

Source: Carpenter, T. M., Carnegie Institution of Washington, Publ. No. 216, Washington, D. C., 1915.

In several conditions there is an apparent

[1] There is definite proof that the tissues of diabetics burn sugar, but at a lower rate than normal tissues.

increase in the RQ due not to changes in the oxidation processes taking place in the body, but to the elimination of CO_2 stored in the alkali reserve. There are two main conditions in which this occurs: (*a*) when hyperventilation eliminates CO_2 in the alveolar air and the blood, without there being an increased formation of CO_2; (*b*) when an excess of acid enters into the blood and displaces CO_2 from bicarbonates; *e.g.*, lactic acid produced by muscular contraction, or ketonic bodies in diabetic ketosis and in acidosis due to the ingestion of ammonium chloride or calcium chloride.

Table 38. Caloric Value of Oxygen in the Body

RQ	*Calories per liter of O_2 consumed*	*Percentage of total heat produced by carbohydrate*	*Percentage of total heat produced by fat*
0.77	4.686	0	100.0
0.75	4.739	15.6	84.4
0.80	4.801	33.4	66.6
0.85	4.862	50.7	49.3
0.90	4.924	67.5	32.5
0.95	4.985	84.0	16.0
1.00	5.047	100.0	0

Source: Zunz and Schumburg, modified by Lusk, and later corrected by Cathcart and Cuthbertson (*J. Physiol.*, **72**, 349, 1931). According to E. P. Poulton (*Proc. Roy. Soc. Med.*, **26**, 1591, 1933), these figures are not completely accurate.

Conversely, an apparent decrease of the RQ can be observed when there is retention of CO_2 without any change in the oxidation processes of the body. Thus CO_2 is retained in hypoventilation and in alkalosis provoked by excess elimination of acids (repeated vomiting, etc.) or by excess ingestion of alkali.

The caloric value of oxygen. Three data are necessary for the calculation of respiratory metabolism: (*a*) the amount of O_2 absorbed, in liters; (*b*) the duration of the experiment; (*c*) the caloric value of 1 liter of oxygen.

The amount of heat produced by 1 liter of O_2 varies according to the substance oxidized (Table 38). One liter of oxygen oxidizing carbohydrate sets free 5.047 kg.-cal., while it gives only 4.686 cal. when fat is oxidized.

The RQ must therefore be known to obtain accurately the caloric value of the oxygen absorbed (Table 38). If the RQ is not known, as is the case when using the Benedict-Roth spirometer, an average caloric value of 4.825 kg.-cal., corresponding to an RQ of 0.82,[1] is attributed to each liter of O_2. If the subject's RQ is above or below 0.82, an error will be committed, but it is never greater than ± 7 per cent.

If, instead of measuring the O_2 absorbed, only the CO_2 eliminated is determined and a caloric value corresponding to an RQ intermediate between 0.7 and 1 is attributed to the CO_2, the maximum possible error will be ± 15 per cent. Moreover, as was explained previously, the elimination of CO_2 is subject to variations that do not correspond to changes in the oxidizations taking place in the body. For these reasons respiratory calorimetry is not based on CO_2 determinations alone, whenever the O_2 consumption can be measured.

In more precise experiments the nonprotein RQ is determined. For example, a subject eliminates 13.5 liters of CO_2, absorbs 16 liters of O_2, and eliminates 0.5 gm. of N in the urine, per hr. For each gram of N produced in the combustion of proteins, there is an output of 4.76 liters of CO_2 and 5.94 liters of oxygen are absorbed. Therefore, in the example given, $0.5 \times 4.76 = 2.38$ liters of CO_2 and $0.5 \times 5.94 = 2.97$ liters of O_2 correspond to protein metabolism; $13.5 - 2.38 = 11.12$ liters of CO_2 and $16 - 2.97 = 13.03$ liters of O_2 correspond to carbohydrate and fat burned. Therefore,

$$\text{Nonprotein RQ} = \frac{11.12 \text{ liters } CO_2}{13.03 \text{ liters } O_2} = 0.85$$

The caloric value of O_2 at this RQ is 4.862; therefore the heat output per hour is $13.03 \times 4.682 = 63.3$ kg.-cal., 50.7 per cent from the oxidation of carbohydrate and 49.3 per cent from that of fat. Protein metabolism gives rise to $2.97 \times 4.485 = 13.3$ kg.-cal. This figure can also be obtained by multiplying the amount of nitrogen excreted by its caloric equivalent, *i.e.*, $0.5 \times 26.51 = 13.3$ kg.-cal. Total heat production per hour in the subject taken as an example is $63.3 + 13.3 = 76.6$ kg.-cal.; 42 per cent of this total is obtained from carbohydrate, 41 per cent from fat, and 17 per cent from protein. A less precise but fairly accurate calculation can be made by considering the RQ as the result of carbohydrate and fat combustion, without taking into account the protein metabolized. Thus nitrogen determinations are not necessary and no considerable error is made.

[1] This is the RQ most commonly observed in fasting subjects, the condition in which basal metabolism is determined.

BASAL METABOLISM

Significance. The total energy converted by the human organism can be differentiated into two parts:

1. A constant or basal metabolism; *i.e.*, the minimum energy necessary for the maintenance of normal function.
2. An additional variable quota of energy expended in work, in the maintenance of the body temperature, and in the postabsorptive condition (a short period following the absorption of food).

The true minimum metabolism is not easy to determine, because what is usually called "basal metabolism" diminishes during sleep, in poorly fed subjects and in sedentary ones. Therefore it is more convenient to define basal metabolism as the energy output observed under certain special conditions, known as "basal conditions," in which exercise, outside temperature, and the recent absorption of food have no influence on energy exchanges.

There are three fundamental basal conditions: (*a*) the subject must be resting comfortably, relaxed, without moving, for ½ to 1 hr. before the determination is made, and must not have performed any violent exercise recently; (*b*) outside temperature must be approximately 20°C. (16 to 25°C.); (*c*) the subject must not have ingested any food for 12 to 16 hr. (postabsorptive period) before the determination. Krogh recommends that no protein be taken on the night before and that the test be performed between 8:00 and 9:00 A.M. The subject should not be in a condition of emotional excitement, but should be quiet, his pulse and breathing regular, and his body temperature normal.

The term "basal metabolism" or "basal metabolic rate" has come into common use, but "standard metabolism" proposed by Krogh would be more appropriate.[1] It is the sum of the metabolic activity of the different organs, *i.e.*, of the heart (4 to 5 per cent), kidneys (5 to 7 per cent), liver, digestive tract, contraction of the respiratory muscles, maintenance of muscle tone, etc. This activity is controlled by the nervous system and the endocrine glands. It is usual to express the basal metabolic rate in calories per square meter per hour.

Basal metabolic rate and body mass. Basal metabolic rate per kilogram of body weight is greater in small animals than in large. If the basal metabolic rate is calculated per square meter of body surface, however, approximately the same figures are found for large and small animals. This fact led to the statement that the basal metabolic rate is proportional to the area of the body surface, known as the "law of surface area," put forward by Rubner (1883) and Richet (1889). This does not mean that heat production is dependent on the heat lost through the skin, because if this were so heat production would cease when the outside temperature was equal to the body temperature, and it does not cease even when the temperature of the environment is higher than the body temperature (Du Bois).

Meeh determined the surface area using the following formula:

$$S = KW^{0.66}$$

where W is the body weight in kilograms and K is a constant which varies for each species but which is approximately equal to 10. Du Bois and Du Bois studied the body surface of man by applying paper patterns on the skin and measuring the total area. They arrived thus at the following empiric, but more accurate, formula:

$$\text{Surface (in sq. cm.)} = \text{Weight (in kg.)}^{0.425} \times \text{Height (in cm.)}^{0.725} \times 71.84$$

By the application of this formula, charts have been made for calculating the body surface of an individual (Figs. 187 and 188) from his body weight and height. These charts are in common use in medical practice.

The "law of surface area" is not strictly accurate because heat production per square meter of body surface is somewhat greater in large animals than in small ones (Benedict); it also varies with the climate (Galvao). Certain authors have discarded this "law" and maintain that heat production is dependent on the active protoplasmic mass (Benedict), on the physiologic (not the physical) weight (Brody), or on a principle of biological similarity (Kayser)—terms which are not easily defined.

The basal metabolism of many homoiothermic

[1] Other terms have been used, such as "fundamental metabolism" (Grundumsatz, Magnus Levy), "postabsorptive metabolism" (Benedict), "minimum physiological energy" (Lefevre), "maintenance metabolism," etc.

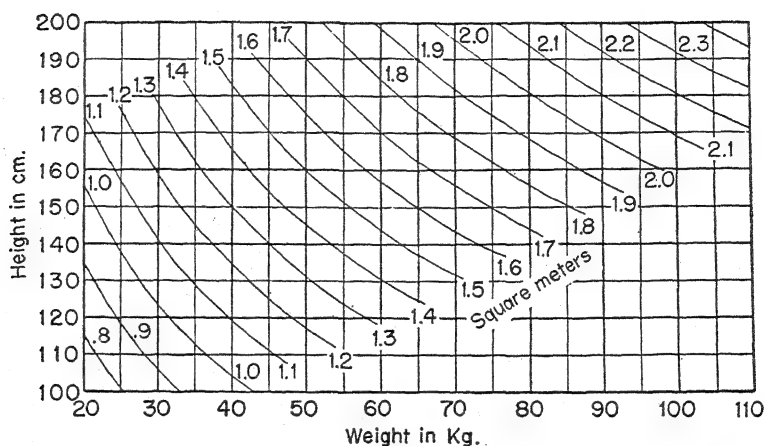



FIG. 187. Du Bois's chart for determining surface area of man from weight and height.

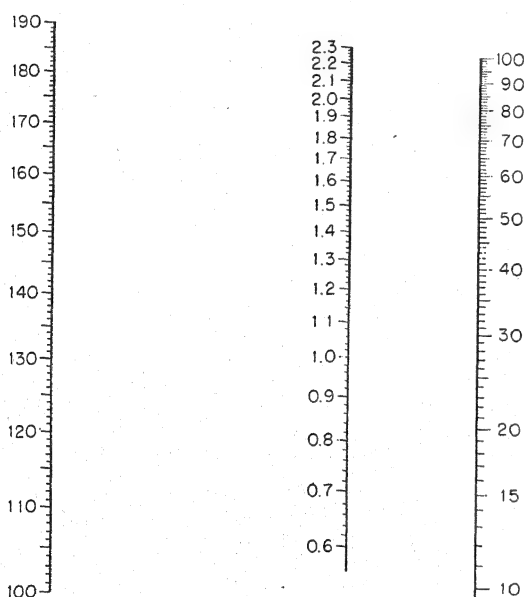



FIG. 188. Boothby and Sandiford's nomogram for determining surface area of man. The point on the scale on the left corresponding to the subject's height in centimeters is joined by a straight line to the point on the scale on the right corresponding to his weight in kilograms. The point where this line crosses the scale in the middle gives the surface area in square meters.

species is proportional to a power of the body weight. According to Brody this power is 0.73, and according to Kleiber it is 0.75. Kleiber (1947) has proposed the following formula for man: BMR = 70 kg.-cal. $\times W^{0.75}$ per day, or BMR = 3 kg.-cal. $\times W^{3/4}$ per hr. Günther[1] gives a similar formula: $\frac{\text{kg.-cal./hr.}}{W^{0.734}} = 3$ (men) or 2.85 (women).

In the city of São Paulo (Brazil) on the Southern tropic, Galvao (1951) found the following figures for men and women of different build:

Lean: $\text{BMR} = \frac{\text{kg.-cal./hr.}}{W^{0.83}} = 2.038$ (men), 2.012 (women)

Medium: $\text{BMR} = \frac{\text{kg.-cal./hr.}}{W^{0.78}} = 2.334$ (men), 2.168 (women)

Obese: $\text{BMR} = \frac{\text{kg.-cal./hr.}}{W^{0.71}} = 3.158$ (men), 2.919 (women)

According to Galvao, in cold and temperate climates, heat production and loss are apparently related to the body surface, but in tropical climates they are related to the metabolically active part of the body weight.

The metabolism of poikilothermic vertebrates is also related to a power of the body weight; at 20°C., BMR = $1.78W^{0.728}$ (Kayser, 1951). Many other anatomical and functional characteristics of mammals are related to a power of the body weight, as if they were conditioned by similar regulatory mechanisms or equilibriums (Adolph, 1949).

The expression of the results. The basal metabolic rate is usually given in kilogram-calories per square meter per hour. Average values in man have been established by several workers (Table 39); 68 per cent of the subjects

[1] GÜNTHER, B., *Bol. Soc. biol. Concepción, Chile,* **18,** 45, 1944; *Medicina, Buenos Aires,* **6,** 261, 1946.

the mean; 90 per cent are within ±15 per cent of the mean, and 95 per cent within ±18 per cent (Fig. 189). In medical practice, values within ±15 per cent of mean are considered normal, and those ±20 per cent are considered as almost certainly due to some abnormal condition. The BMR is usually expressed as the percentage deviation from the normal corresponding to the age and sex of the subject. For example, a man twenty-five years old has an output of 45 kg.-cal. per sq. m. per hr., the mean for that age and sex being 40 kg.-cal. per sq. m. per hr. The BMR will be expressed as follows:

$$\frac{(45 - 40)100}{40} = +12.5 \text{ per cent}$$

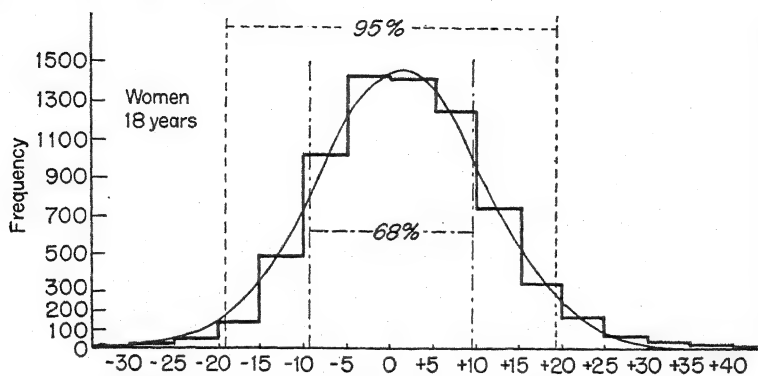



FIG. 189. Normal basal metabolism. Frequency curve of normal values observed in 7,117 women at the Mayo Clinic. The average found in this series was +1.94 with a standard deviation of 9.7. (*Boothby, W. M., J. Berkson, and W. A. Plummer, Tr. Am. A. Study of Goiter, 1937.*)

It is advisable to repeat the determination several times, to arrive at the BMR corresponding to a given subject, especially if abnormal results are obtained in the first observation. In this way mistakes due to emotional disturbance or infringement of one of the basal conditions are avoided.

Table 39. Standard Basal Metabolic Rate

Age, ears	*Average kg.-cal. per hr. per sq. m. of body surface*		*Age, years*	*Average kg.-cal. per hr. per sq. m. of body surface*	
	Males	*Females*		*Males*	*Females*
5	53.0	51.6	20–24	41.0	36.9
6	52.7	50.7	25–29	40.3	36.6
7	52.0	49.3			
8	51.2	48.1	30–34	39.8	36.2
9	50.4	46.9	35–39	39.2	35.8
10	49.5	45.8	40–44	38.3	35.3
11	48.6	44.6	45–49	37.8	35.0
12	47.8	43.4			
13	47.1	42.0	50–54	37.2	34.5
14	46.2	41.0	55–59	36.6	34.1
15	45.3	39.6	60–64	36.0	33.8
16	44.7	38.5	65–69	35.5	33.4
17	43.7	37.4			
18	42.9	37.3	70–74	34.8	32.8
19	42.1	37.2	75–79	34.2	32.3

Source: BOOTHBY, W. M., J. BERKSON, *et al.*, *Am. J. Physiol.*, **116**, 468, 485, 1936; **121**, 669, 1938.

FACTORS THAT MODIFY BASAL METABOLISM

The basal metabolic rate varies with age, sex, size, the nutritional condition (overfeeding, malnutrition), athletic training, sleep, climate, altitude, and body temperature. It is constant in the same subject on different days.

Age. The newborn child has a low BMR (26 kg.-cal. per sq.m. per hr.). It rises gradually with age up to a maximum when the child is between three and six years old; then it falls until there is a slight rise at puberty, followed by a gradual decrease up to about twenty years of age. Then it falls at a slower rate as age advances (Table 39, Fig. 190).

Sex. The BMR is approximately 7 per cent higher in the male than in the female of the same age.

Malnutrition. This condition is perhaps the most frequent cause of low BMR in otherwise normal subjects.

Training. Athletes in training may have a BMR 3 to 6 per cent above the average, but this depends in great part on the capacity of the subject to relax his muscles while the determination is being made. Subjects with sedentary habits may have a low BMR.

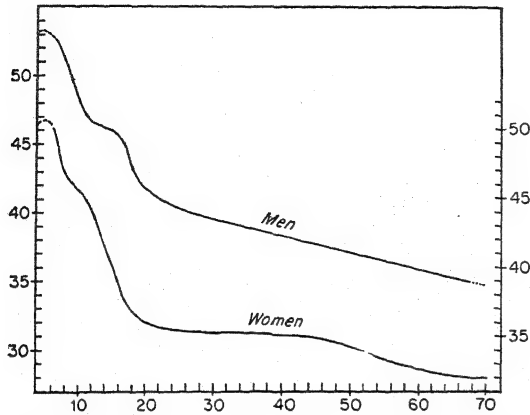



FIG. 190. Influence of age and sex on basal metabolic rate. Abscissa, age in years; ordinate, BMR in kilogram-calories per square meter per hour.

Sleep. During deep sleep the BMR falls from 6 to 13 per cent.

Climate and race. Ozorio de Almeida[1] found a BMR on an average 20 per cent lower in Rio de Janeiro than the normal figures for countries with a temperate climate, and he maintained that living in a tropical climate causes a decrease in BMR. Several observers have found low basal metabolic rates in subjects living in hot climates in Australia, Brazil, India, China, and the southern United States, but not as low as those reported by Ozorio de Almeida (Castro, Oliveira, Orsini, Galvao).[2] Normal basal metabolic rates have been found in Argentina, Cuba (Coro), Peru (Hurtado), and Mexico (Roca). Racial differences apparently also have an influence on the BMR; thus the Mayas of Yucatan (Benedict) and the Araucarian Mapuche Indians of southern Chile (Pi-Suñer) have a slightly higher average BMR than whites living in a moderate climate. It is sometimes difficult to distinguish between the effects of climate (heat and humidity), race, nutritional condition, work, etc.

Altitude. There is little change in the BMR unless the oxygen partial pressure falls to nearly half the oxygen pressure at sea level; with lower oxygen tensions, the BMR increases. In severe anoxia the BMR falls gradually.

Psychic influences. Mental work, even when it is intense, does not modify the BMR significantly; sometimes there is a slight rise of not more than 4 per cent. In emotional states there is a rise of 5 to 20 per cent due to increase in muscular tone, tachycardia, and increased respiratory frequency.

The endocrine glands. *Thyroid.* The thyroid gland plays an important part in the control of oxidations in the organism, therefore of the BMR. Thyroid secretion has a continuous stimulating effect on basal metabolism. Thyroid insufficiency causes a decrease in the BMR of 25 to 40 per cent (Fig. 191), but a rate 15 to 20 per cent below the average is not usually due to hypothyroidism but to other causes (Boothby). Thyroid given by mouth or thyroxin injections raise the BMR of hypothyroid patients. Hyperthyroidism due to over-activity of the thyroid or provoked experimentally by excess thyroid administration causes the BMR to rise. Thyroidectomy in hyperthyroid patients produces a fall to normal or below normal. BMR determinations are of great importance for the diagnosis of the level of thyroid function and in the control of treatment of hyperthyroidism and hypothyroidism.

Hypophysis. The anterior lobe of the hypophysis stimulates the development of thyroid function and maintains its normal level. Anterior hypophyseal lobe insufficiency causes hypothyroidism and therefore a fall in the BMR. Hyperfunction of the anterior hypophyseal lobe, on the contrary, causes hyperthyroidism and a rise in the BMR.

Adrenals. In uncompensated, but not in compensated, adrenal insufficiency there is a low BMR. A sudden discharge of adrenaline from the adrenal medulla causes a rise of the BMR during several minutes, or at most a few hours.

Gonads. The difference in the BMR in the sexes has already been mentioned. Castration diminishes basal metabolism in 20 to 25 per cent of human subjects. Menstruation has little or no effect; frequently there is a slight premenstrual rise in temperature and in BMR and a decrease during the flow.

Pregnancy. During the first half of pregnancy there is no change in the BMR, but in the latter half there is a rise of about 8 per cent at the fifth month and 14 to 33 per cent during the last

[1] OZORIO DE ALMEIDA, A., *J. de physiol. et de path. gén.*, **18**, 713, 1919; **18**, 958, 1920.

[2] MOURA CAMPOS, F., Atti VII Conv. Volta, Roma, 1937, p. 68; *Folha med.*, **20**, 73, 1939.

2 months. The total basal metabolism is equal to the BMR of the mother plus that of the fetus. There is no proof that maternal metabolism is stimulated by pregnancy.

Other factors. Occupation, race, diet, work performed or emotional stress on the day before the determination, the effect of novelty in the first determination, and the season of the year are factors that sometimes have a slight influence on the BMR.

Disease. The BMR is increased in hyperthyroidism (25 to 80 per cent), fever (13 per cent per degree centigrade), dyspnea (25 to 60 per cent), leukemia (20 to 80 per cent), polycythemia (10 to 40 per cent), and some cases of severe anemia. It is low in hypothyroidism (30 to 40 per cent), hypophyseal insufficiency, adrenal insufficiency (Addison's disease), and malnutrition.

Drugs. The following drugs increase the BMR: adrenaline, thyroid and thyroxine, caffeine, benzedrine, dinitrophenol and dinitrocresol, etc. The BMR is depressed by anesthetics and by depression of the central nervous system and muscle tone.

Fig. 191. Basal metabolic rate in thyroid disease. From left to right: spontaneous myxedema; normal distribution (smooth curve); adenomatous goiter, with and without clinical hyperthyroidism; exophthalmic goiter. Ordinate, percentage frequency for each group. Abscissa, metabolic rate percentage deviation from normal (= 0). Almost all cases of myxedema are below normal, and those of exophthalmic goiter are above normal; cases with thyroid adenoma form a heterogeneous group. (*Boothby, W. M., J. Berkson, and W. A. Plummer, Ann. Int. Med., vol. 11, p. 1014, 1937.*)

The Effect of Exercise, Temperature, and Food on Total Metabolism

The simple act of standing up and performing a few movements causes heat production to rise by 25 to 50 per cent; in the course of normal activity the metabolic rate is frequently twice the basal level, and in violent exercise it can be 10 to 16 times the basal rate (see Chap. 47).

External temperature has a marked influence on heat production. A minimum BMR is observed when the subject is at a temperature of 18 to 30°C. (20 to 25°C.). Environmental temperature above 30°C. causes a slight rise in the metabolic rate and in body temperature. When the temperature falls below 15°C. muscular tone increases, there may be shivering, and heat production rises (see Chap. 48). In fever there is a rise in heat production of approximately 13 per cent of the BMR for each degree centigrade increase (Du Bois).

Metabolism is stimulated by the absorption of food, especially of protein. This is known as the specific dynamic action of foodstuffs (see Chap. 43, Protein Metabolism).

BIBLIOGRAPHY

Adolph, E. F., Quantitative Relations in the Physiological Constitution of Mammals, *Science*, **109**, 579, 1949.

Brody, S., "Bioenergetics of Growth," Reinhold, New York, 1945.

Du Bois, E. F., "Basal Metabolism in Health and Disease," 3d ed., Lea & Febiger, Philadelphia, 1936.

Galvao, P. E., Metabolismo Basal nos Tropicos, *Ciencia e cultura*, **3**, 18, 1951; *J. Appl. Physiol.*, **1**, 383, 395, 1948.

Kayser, C., La Loi des surfaces, *Rev. scient.*, **89**, 267, 1951.

Kleiber, M., Body Size and Metabolic Rate, *Ann. Rev. Physiol.*, **6**, 123, 1944; *Physiol. Rev.*, **27**, 511, 1947.

Lang, K., and O. F. Ranke, "Stoffwechsel und Ernährung," Springer, Berlin, 1950.

Lusk, G., "The Science of Nutrition," 4th ed., Saunders, Philadelphia, 1928.

Talbot, F. B., Basal Metabolism of Children, *Physiol. Rev.*, **5**, 477, 1925.

Terroine, E. F., and E. Zunz, "Le Métabolisme d base," Presses Universitaires de France, Paris, 1925.

CHAPTER 41

Carbohydrate Metabolism

Absorption. The process of digestion in man splits carbohydrates into monosaccharides, and they are absorbed as such, mainly in the form of the hexoses: glucose, fructose, and galactose. Absorption takes place principally in the small intestine, and only in small amounts in the colon. Insignificant quantities of carbohydrate are absorbed in the stomach, except when concentrated solutions are given for experimental purposes. Sugars pass through the intestinal mucosa exclusively in a one-way direction, from the lumen of the intestine to the blood stream. The velocity of absorption varies from one sugar to another. If the conventional value of 100 is given to the amount of glucose absorbed in unit time, the following values are found in the rat for the absorption of other sugars: galactose, 110; fructose, 43; mannose, 19; xylose, 15; arabinose, 9 (Cori). These figures show that glucose and galactose are absorbed at a much higher rate than any of the others. This selective absorption is impaired if the animals are kept at low temperatures (from 0 to 20°C.), in some endocrine insufficiencies (adrenal, thyroid, hypophysis), and in cases of intoxication by certain drugs. In these cases all sugars are absorbed at the same rate. The mechanism responsible for selective absorption has not yet been clearly demonstrated; nevertheless many facts indicate that it is due to phosphorylation of glucose and galactose, according to Verzár's hypothesis, while other sugars are absorbed by simple diffusion. Monoiodoacetic acid and phlorhizin inhibit phosphorylation, thus suppressing selective absorption.

The thyroid plays a definite part in the selective absorption of glucose and especially of galactose. These sugars are absorbed at a much slower rate in hypothyroidism and at a higher rate in hyperthyroidism. In adrenal insufficiency the selective absorption of glucose is disturbed. The normal condition is restored by treatment with corticoadrenal hormones or simply by giving sodium chloride.

Distribution in the organism. Glucose and glycogen are the principal carbohydrates in the human body; milk contains lactose. Carbohydrates are also found combined with fats (galactosides or cerebrosides), proteins (glycoproteins), or phosphoric acid (hexosephosphates and triosephosphates). Desoxyribose and ribose are found in nucleic acids and their derivatives, and in certain coenzymes. Phosphoric esters of glucose and fructose are of importance in the intermediate metabolism of carbohydrates.

Glycogen. This carbohydrate, discovered by Claude Bernard, is a polysaccharide corresponding to the formula $(C_6H_{10}O_5)^x$. It is a white powder, giving opalescent colloidal solutions, and taking on a dark red color in the presence of iodine. It is hydrolyzed by acids and amylolytic enzymes, giving glucose as a final product.

Glycogen[1] is found in many tissues, and there are large quantities in the liver, muscle, and heart. Glycogen and glucose concentrations vary in the different tissues (Table 40).

Although the liver has a higher glycogen concentration than muscle, the total amount of liver glycogen (75 to 200 gm.) is usually less than that of muscle glycogen (180 to 300 gm.), because the liver makes up only 3.3 per cent of the body weight, while the muscles are almost 50 per cent. Chemical transformation of glycogen varies considerably

[1] Glycogen is usually isolated and estimated by Pflüger's method or a modification of this method. The tissue is boiled for 2 to 4 hr. in 60 per cent KOH, which dissolves protein but does not attack glycogen; alcohol up to 70 per cent is then added to precipitate glycogen, which is washed and hydrolyzed with 2.2 per cent HCl for 2 hr. Glucose thus formed is then estimated.

in the different tissues, because of the different chemical systems responsible for its formation and destruction.[1] There are remarkable differences between liver and muscle glycogen (Table 41). Every day 69 per cent of the glycogen in the liver is renewed (Stetten). The process of

Table 40. Distribution of Carbohydrate in the Organism

Carbohydrate	*Liver, gm. per 100 gm.*	*Muscle, gm. per 100 gm.*	*Blood, gm. per 100 cc.*
Glycogen......	0.2–10*	0.4–1.0†	Traces
Glucose.......	0.05–0.15	0.02–0.04	0.08–0.12
Lactic acid....	0.01	0.01	0.01

* A maximum value of 18 gm. per cent has been reported.

† A maximum value of 3.5 gm. per cent has been reported.

glycogen formation is known as glycogenesis, and its disintegration is called glycogenolysis.

Hexosamines. Glucoseamine and galactoseamine are found in the body; in the former, one OH is replaced by an amine (NH_2). There are also uronic acids, the most common one being glucuronic acid; its formula is the same as that of glucose, but a carboxyl group substitutes for the primary alcohol. Condroitinsulfuric acid is an important constituent of cartilage and tendons; it is a polymer of glucuronic acid and glucosamine with sulfate and acetyl groups.

Table 41. Behavior of Liver and Muscle Glycogen

	Liver glycogen	*Muscle glycogen*
Concentration......	Varies considerably	Varies little
Part played in blood-sugar regulation..	Important	None
Transformed to....	Glucose	Lactic acid

Hyaluronic acid, found in the umbilical cord, the synovial fluid, the vitreous body, and the skin, is made up of glucuronic acid and acetylglucosamine. An enzyme, hyaluronidase, hydrolyzes hyaluronic acid and diminishes its viscosity. If hyaluronidase is added to a fluid injected subcutaneously, this diffuses rapidly in the dermis (spreading factor). Certain anaerobic bacteria and snake venoms spread quickly because they contain this enzyme. The testes have a large amount of hyaluronidase, which is excreted in the seminal fluid and aids the penetration of the spermatozoon into the ovum by

Intestine
Blood
5 gm.
Kidney
Tissue fluid
12 gm.
Muscle glycogen 300 gm.
Hepatic glycogen 180 gm.
Fat
Oxidation CO_2 + H_2O

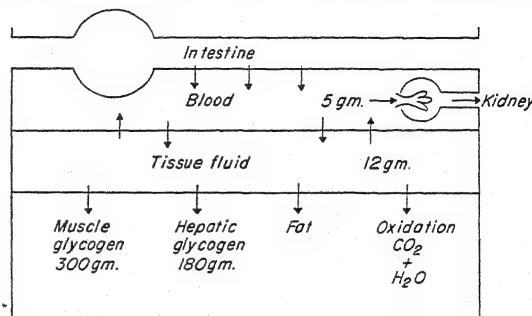



FIG. 192. Diagram illustrating the fate of carbohydrate in the body.

dissolving the hyaluronic acid cuticle which surrounds it.

Origin and fate of carbohydrate in the organism. Carbohydrate in the body arises (*a*) from carbohydrate in foodstuffs, and (*b*) from substances other than carbohydrate (glyconeogenesis) such as certain amino acids and in smaller amounts from fatty acids.

Glucose is absorbed from the intestine into the blood stream; it then passes by diffusion through the capillary walls into the tissue fluids (Fig. 192). Glucose then follows one of three main paths: (*a*) conversion into glycogen; (*b*) conversion into fat; (*c*) oxidation into CO_2 and H_2O, which are eliminated.[1]

Up to quite recently it was believed that glucose absorbed from the intestine was mainly converted into glycogen. Experiments performed by feeding rats with "labeled" glucose (*i.e.*, glucose containing deuterium, or "heavy" hydrogen, in its molecule) have shown that only 3 per cent is deposited as glycogen and 30 per cent is converted into fat.[2]

INTERMEDIARY CARBOHYDRATE METABOLISM

The synthesis of glycogen and the oxidation of glucose into CO_2 and H_2O are performed only after a process of esterification with phosphate. The different reactions are illustrated on

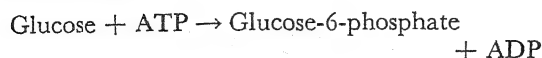
[1] Thus in severe diabetes glycogen diminishes considerably in the liver and to some extent in muscle; on the contrary it increases in the heart, leukocytes, and kidneys.

[1] Part of the CO_2 can transitorily form part of intermediate metabolic products.

[2] STETTEN, DE W., *J. A. M. A.*, **132**, 373, 1946; *Proc. Am. Diabetes A.*, **7**, 67, 1948.

page 403. All these reactions can be considered as reversible (the arrows have been drawn in only one direction in order to simplify the diagram).

Free glucose is not used until it is phosphorylated by an enzyme, hexokinase, which catalyzes the following reaction:

Glucose + ATP → Glucose-6-phosphate + ADP

ATP signifies adenosinetriphosphate, and ADP adenosinediphosphate.

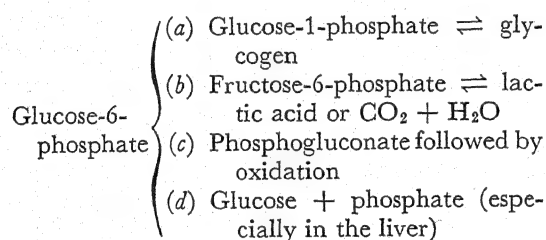
In stage 1, phosphate is transferred from ATP to glucose, which must be phosphorylated in order to be utilized. In stage 2, phosphis ate transferred from position 6 in glucose-1-6-diphosphate to position 1 of glucose-6-phosphate,[1] with the formation of glucose-1-phosphate and regeneration of glucose-1-6-diphosphate; which has therefore acted catalytically. In stage 3, phosphorylase converts glucose-1-phosphate into glycogen and releases inorganic phosphate.

```
  N=C·NH2         ┌──────O──────┐
  |   |           |  OH    OH   |                    O       O       O
  |   |           |   |     |   |                    ‖       ‖       ‖
  HC  C—N———————C—C———C—C—CH2—O—P—O—P—O—P—OH
  ‖   ‖   \       |   |     |   |                    |       |       |
  |   |    CH     H   H     H   H                    OH      OH      OH
  |   |   //
  N—C—N
```

Adenosinetriphosphate.

In the tissues adenosinemonophosphate is also found. These adenosinephosphates act as phosphate carriers—donors or acceptors, according to the reaction. Further on, processes will be described in which ADP accepts phosphate and is converted into ATP.

After glucose-6-phosphate has been formed by the action of hexokinase, several metabolic paths are opened:

Glucose-6-phosphate:
- (*a*) Glucose-1-phosphate ⇌ glycogen
- (*b*) Fructose-6-phosphate ⇌ lactic acid or $CO_2 + H_2O$
- (*c*) Phosphogluconate followed by oxidation
- (*d*) Glucose + phosphate (especially in the liver)

It should be noted that other hexoses such as fructose, mannose, and galactose are also phosphorylated by ATP.

Glycogen synthesis and breakdown. The conversion of glucose into glycogen takes place in three stages:

1. Glucose + ATP → glucose-6-phosphate + ADP (*hexokinase*)
2. Glucose-6-phosphate → glucose-1-phosphate (*phosphoglucomutase*, glucose-1-6-diphosphate acts as coenzyme)
3. Glucose-1-phosphate → glycogen + phosphate (*phosphorylase*)

Conversion of glycogen into glucose follows the same path in the opposite direction. Glycogen, as a result of the action of phosphorylase, gives glucose-1-phosphate, which is converted by phosphoglucomutase into glucose-6-phosphate, and finally free glucose and phosphate are formed.

Glycogen is made up of many units of glucose, united like the links of a chain. An oxygen atom unites the first carbon of one glucose molecule to the fourth carbon of the next (see formula on page 403). This is known as "1-4 glucosidic linkage." There are also 1-6 linkages.

Phosphorylase acts on a terminal 1-4 linkage of the glycogen molecule, and an inorganic phosphate is joined to it, setting free glucose-1-phosphate, thus shortening the glycogen chain by one glucose unit. This is a reversible process, and according to the concentration of glucose-1-phosphate, the glycogen chain will be lengthened or shortened, *i.e.*, glycogen will be synthesized or broken down.

Phosphorylase catalyzes only the formation or breaking of 1-4 linkages; another, so far little-known, enzyme acts on 1-6 linkages of glycogen.

The reversible transformation of glycogen ⇌ glucose that takes place in the process of carbohydrate metabolism is due to phosphoryla-

[1] Leloir, R. F., R. E. Trucco, C. E. Cardini, A. Paladini, and R. Caputto, *Arch. Biochem.*, **19**, 339, 1948; **22**, 87, 1949.

tion; glycogen, however, can also be broken down by amylase, which breaks the glucosidic linkages by adding water instead of phosphate.

The formation of pyruvic and lactic acids. Glucose-6-phosphate not only may be converted into glycogen, but it may also follow a catabolic path which leads by a series of reactions to the formation of pyruvic acid. In this series fructose diphosphate is formed by phosphorylation with ATP. This molecule is then split into two molecules of triosephosphate, and adenosinephosphate is charged twice, by reacting with diphosphoglyceric acid and with phosphopyruvic acid. ATP is in this manner resynthesized, and it can then take part in the phosphorylation of another glucose molecule or in other reactions.

Di-phosphopyridin nucleotide (DPN) or coenzyme I (see Chap. 34) also takes part in these processes. This substance can be present in the oxidized or reduced form. The latter ($DPNH_2$) has two more hydrogen atoms than the former. In the oxidation of phosphoglyceraldehyde, coenzyme is reduced, DPN → $DPNH_2$, and in the reduction of pyruvic acid to lactic acid the reverse occurs, $DPNH_2$ → DPN.

Direct oxidation of glucose-6-phosphate. A less well-known process undergone by glucose-6-phosphate is its oxidation into phosphogluconic acid, which is converted into ribose-phosphate.

The formation of glucose. A very active phosphatase has been found, especially in the liver, which hydrolyzes glucose-6-phosphate into glucose and inorganic phosphate. Three enzymes take part in the conversion of glycogen into glucose in the liver: phosphorylase, phosphoglucomutase, and phosphatase. Amylase perhaps also takes some part in this process.

The role of adenosinephosphate. Muscular contraction is produced by the interaction of adenosine-

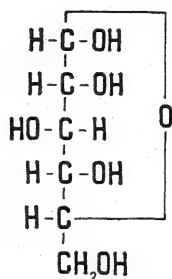
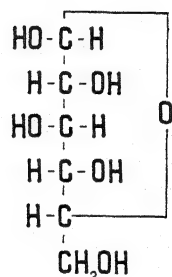
triphosphate (ATP) and actomyosin, one of the proteins in muscle. Chemical energy set free from ATP is converted into mechanical energy, the muscle contracts, and ATP loses some of its phosphate. Probably ATP acts as a source of energy in many other physiologic processes; therefore the origin of this substance is of great interest.

Adenosinephosphate is found in the tissues in three degrees of phosphorylation, *i.e.*, as adenosinemonophosphate (AMP), as adenosinediphosphate (ADP), and as adenosintriphosphate (ATP). When AMP accepts one phosphate group it is converted into ADP, which on accepting a second phosphate becomes ATP. This charging of adenosinephosphate cannot be done with inorganic phosphate; it is realized by the interchange of phosphate combined with other substances. Several phosphorylation reactions take place without adenosinephosphate (see diagram on page 403). Conversion of glycogen to glucose-1-phosphate and oxidation of phosphoglyceraldehyde to diphosphoglyceric acid are carried out by the direct introduction of inorganic phosphate. In anaerobic conversion of glycogen to lactic acid, adenosinephosphate acts as a phosphate donor in the production of fructosephosphate and as phosphate acceptor in reactions with 1–3 diphosphoglyceric acid and with phosphopyruvic acid. As each hexose molecule gives rise to two triose molecules for each glucose unit derived from glycogen, adenosinephosphate is charged three times with phosphate (two phosphates are obtained from diphosphoglyceric acid and two from phosphopyruvic acid, and one is lost in the phosphorylation of fructose phosphate). If the reaction begins with free glucose instead of glycogen, only two charges of phosphate are available, because one is lost in the initial phosphorylation of glucose.

Oxidation of glucose in oxygen gives rise to a considerably greater process of phosphorylation; there are approximately 36 charges of phosphate for each glucose molecule oxidized. The intimate mechanism of aerobic phosphorylation is not yet clearly understood, but undoubtedly it is a much more efficient process than anaerobic phosphorylation. This greater efficiency is also evident in thermodynamic studies; the oxidation of one glucose molecule sets free 680 kg.-cal., while its conversion to lactic acid gives rise to only 54 kg.-cal.

THE BLOOD SUGAR

Distribution of glucose in the body fluids. Glucose is dissolved in the water of all extracellular body fluids in approximately the same concentration, because of its great diffusibility. If a subject has 1 gm. of glucose per liter of blood plasma, there will be approximately 20 gm. in all the extracellular fluids, of which 5 gm. will be in the blood plasma and 15 gm. in the tissue fluids; the latter is therefore the principal store of glucose as such.

The sugar in the blood is D-glucose; on dissolving, an equilibrium is rapidly established between two isomers α- and β-glucose, which have been obtained in the pure form.

α-d-GLUCOSE	β-d-GLUCOSE
H-C-OH	HO-C-H
H-C-OH	H-C-OH
HO-C-H (O)	HO-C-H (O)
H-C-OH	H-C-OH
H-C	H-C
CH_2OH	CH_2OH
$[\alpha]D = 111°$	$[\alpha]D = 19 \cdot 3°$
MP = 146°	MP = 148–150°

EQUILIBRIUM AT

$[\alpha]D = 52 \cdot 5°$

Glucose is in a free and diffusible form in the blood; this has been demonstrated by ultrafiltration *in vivo* (Brull) and by the vividiffusion method of Abel. An even more decisive proof is given by the fact that the glucose concentration in glomerular fluid is the same as that in blood plasma (Richards, Walker; see Chap. 63).

Acid hydrolysis of blood plasma causes the formation of the so-called "protein sugar," called "galactoglucosamine-mannose" (Bierry).[1] The metabolic significance of this substance is as yet unknown. Glucose is the principal sugar found in blood plasma; there are only very small amounts of other sugars (fructose). In rare abnormal cases fructose and lactose appear in the urine; the appearance of the latter is related to the function of the mammary gland. Pentoses are sometimes found in the urine. In blood plasma a small amount of a volatile reduc-

[1] BIERRY, H., and F. RATHERY, "Introduction à la physiologie des sucres," J. C. Baillière et fils, Paris, 1835.

ing substance has been found; it is apparently acetaldehyde.

In the estimation of blood sugar by the usual methods, glucose concentration is calculated by the reducing activity of the blood or plasma filtrates.[1] This gives the apparent glucose concentration; real glucose is less than this, because 20 to 27 per cent of the reducing activity is due to other substances, such as creatinine, thioneine, glutathione, etc.[2] The real concentration of sugar is obtained by measuring the reducing power before and after submitting the plasma to the enzymatic action of brewer's yeast, which destroys glucose without altering other reducing substances. The concentration of glucose is the same in the water of the plasma and the erythrocytes, but the latter contain about 20 per cent less glucose because they have a higher protein concentration, and therefore less water than plasma.

Glucose is rapidly broken down in shed blood (glycolysis), and lactic acid is formed. Glycolysis is due mainly to the blood cells. Therefore, when determining the blood sugar, proteins should be precipitated immediately after the blood is drawn, or else the blood should be dried on filter paper or by anhydrous Na_2SO_4; otherwise part of the glucose will be destroyed. Glycolysis can also be partially inhibited by fluoride or monoiodoacetate.

The fasting blood sugar, as determined by current methods, varies between 70 and 100 mg. per cent. These figures correspond to the reducing activity of the blood or plasma; the real concentration of glucose is only 50 to 80 mg. per cent. This is the normal blood-sugar level. A concentration above the normal maximum (usually fixed at 120 mg. per cent) is called hyperglycemia; concentrations below the normal minimum are known as hypoglycemia.

Sugar-tolerance curves. During the absorption period following digestion the blood sugar rises, but never above 160 mg. per cent in normal subjects. If 50 to 100 gm., or 1 gm. per kg. of body weight, glucose is given by mouth to a normal individual, the blood sugar rises, reaching a maximum level not greater than 170 mg. per cent, in about 30 min.; afterward the blood sugar falls slowly, returning to the initial level in about 2 hr. (Fig. 193). Frequently a drop below the normal is observed after the rise, with a subsequent return to the initial level.

[1] See technical manuals for the techniques used.

[2] Somogyi, M., *J. Biol. Chem.*, *75*, 33, 1927; **78**, 117, 1928. Reducing substances that are not glucose are more abundant in the erythrocytes (about 40 mg. per 100 ml.) than in plasma (10 mg. per 10 ml.). Modern improved techniques for blood-sugar analysis exclude more and more reducing substances that are not glucose.

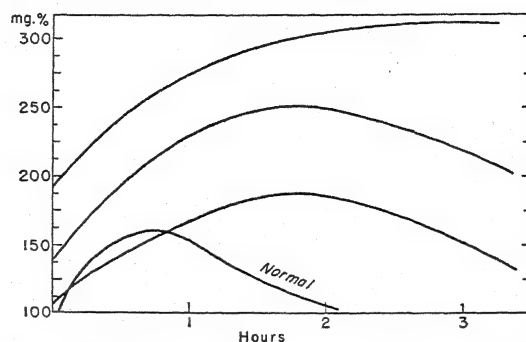



Fig. 193. Blood-sugar curves following the ingestion of glucose (1 gm. per kg. of body weight) in a normal subject and in three cases of diabetes. As the severity of diabetes increases the blood sugar rises to a higher level and remains high for a longer time.

The blood-sugar curve thus obtained is conditioned by the time taken to reach the intestine, the speed of absorption, and the mechanisms that control carbohydrate metabolism. The curve may be prolonged and flattened by retarded gastric evacuation or slow intestinal absorption (thyroid, adrenal, or hypophyseal insufficiency). It is higher and more prolonged in diabetes. The subject is considered to be suffering from diabetes if the blood sugar is above 130 mg. per cent when fasting, and rises to 180 mg. per cent or more ½ hr. after the ingestion of 50 to 100 gm. glucose, without returning to the normal level in 3 hr. In hyperthyroid subjects the blood-sugar curve is higher than the normal after the ingestion of glucose, and especially of galactose (Althausen), owing to the increase in the speed of absorption of these sugars in this condition. For this reason a hyperthyroid subject should not be considered diabetic unless the blood sugar when fasting is 150 mg. per cent or more and rises to 200 mg. per cent after the ingestion of glucose.

The blood-sugar curve following the ingestion of glucose is known as the sugar-tolerance curve. It is conditioned by the previous diet of the subject;[1] the blood sugar rises less in subjects fed a diet rich in carbohydrates, and more in those fed a diet poor in carbohydrates (Fig. 194).

[1] Himsworth, H. P., *Brit. M. J.*, **1**, 719, 1940.

Until recently it was believed the previous diet modified only the response of the liver, but in eviscerated rats the peripheral (extrahepatic) tissues burn preferentially the principal constituent of the previous diet, either fats or carbohydrates.[1]

FIG. 194. Effect of carbohydrate content of previous diet on the peripheral utilization of glucose. Two experiments in the same normal subject; in each experiment 50 gm. glucose was given by mouth. The area between the arterial and venous blood-sugar curves is approximately the same in both experiments; this suggests that in both cases the same amount is removed from the blood by the tissues. [*Himsworth, H. P., Lancet, vol. 55 (2), p 118, 1939.*]

If the dose of glucose is given in two parts, the second dose is not followed by a rise in blood sugar (Hamman and Hirschmann, 1919; Staub and Traugott, 1922).[2] The Exton-Rose[3] test is based on this fact. The subject is given 50 gm. of glucose by mouth, and 30 min. later another 50 gm. The blood sugar does not rise above 160 mg. in the first hour in nondiabetic subjects; in diabetic ones, it rises above 180 mg., and in doubtful cases it rises to 160 to 180 mg. per cent.

The capacity to assimilate glucose has also been measured, but less frequently, by continuous intravenous injection of glucose. In normal subjects a dose of 0.5 to 0.75 gm. per kg. of body weight per hour does not produce glycosuria, while 0.9 gm. per kg. per hr. usually does. When the administration of glucose by mouth is followed by an abnormal tolerance curve and the intravenous injection of glucose gives a normal curve, there is a disturbance in the absorption of carbohydrate, but utilization of sugar (glucose or galactose) is normal.

Ingestion of 150 to 200 gm. glucose seldom provokes glycosuria, and there is never any considerable loss of sugar in the urine even with larger doses (300 to 500 gm.). These tests involving the provocation of alimentary glycosuria are not often performed now, because the determination of the fasting blood sugar and the tolerances curves are of greater value in the examination of carbohydrate metabolism.

The difference in glucose concentration of arterial and venous blood is established by determining glucose in samples obtained at the same time from an artery or by venous puncture. This does not give an accurate idea of the rate of glucose consumption by the tissues, because the speed with which the blood circulates is an important factor in this difference.

Insulin-tolerance curves are obtained by injecting a dose of this hormone and following the changes in blood sugar, noting the degree and duration of hypoglycemia and the time taken to recover the normal level. Himsworth reports that in subjects fed a previous diet with a high carbohydrate content, insulin has a more marked effect than in subjects on a low carbohydrate diet.

[1] ROBERTS, S., Z. T. SAMUELS, and R. M. REINECKE, *Am. J. Physiol.*, **140**, 639, 1944.

[2] TRAUGOTT, K., *Klin. Wchnschr.*, **1**, 892, 1922.

[3] EXTON, W. G., and A. R. ROSE, *Am. J. Clin. Path.*, **4**, 381, 1934.

THE REGULATION OF CARBOHYDRATE METABOLISM

The Regulation of the Blood-sugar Level

The concentration of glucose in blood is an important index of carbohydrate metabolism. The blood-sugar level is dependent on the balance between glucose entering and that leaving the blood. Glucose entering the blood stream (Fig. 192) comes from (*a*) digestive absorption, or subcutaneous or intravenous injection of glucose; (*b*) glucose in the tissue fluids; (*c*) secretion of glucose by the liver. Glucose leaves the blood stream by (*a*) diffusion into the tissue fluids; (*b*) storage as glycogen in the liver, muscles, etc.; (*c*) conversion into fat; (*d*) combustion in the tissues with the formation of CO_2 and water; (*e*) in certain cases, renal excretion.

Digestive absorption of glucose has already been considered. The tissue fluids are a vast store of free glucose in equilibrium with the blood plasma. Absorbed or injected glucose rapidly passes into the tissue fluids, where it is temporarily stored.[1] Part of it is then taken up by the cells, which burn or store it, and part diffuses back into the blood. The tissue fluids also, at times, receive glucose from cells where it has been stored, and from there sugar diffuses into the blood.

The main processes of carbohydrate metabolism can be summarized as follows: The liver secretes glucose into the blood stream, from which it diffuses into the tissue fluids. Glucose is then consumed by the tissues, especially by the muscles, which form 45 per cent of the body weight. The formation and consumption of glucose are controlled by endocrine (pancreatic, anterior hypophyseal, adrenal, etc.) and nervous factors.

[1] Cannon refers to this as storage by flooding, in contrast to storage by segregation, which consists in the conversion of glucose into glycogen.

Fig. 195. Fall in blood sugar following hepatectomy in chloralosed dogs. Average figures of four animals with pancreatic diabetes, five with metathyroid diabetes, five with meta-alloxan diabetes, four with metahypophyseal diabetes, and five normal animals.

The Role of the Liver

The liver is the central organ in carbohydrate metabolism and the main factor in the control of the blood-sugar level. Its principal functions in this respect are (*a*) the secretion of glucose into the blood stream, thus maintaining the blood-sugar level; (*b*) the formation of glycogen within the hepatic cells when there is hyperglycemia, and the breaking down of glycogen when there is hypoglycemia; (*c*) the conversion of fructose, galactose, mannose, and lactic and pyruvic acids into glycogen, which is broken down and secreted as glucose when needed, whatever the substances used in its formation; (*d*) glyconeogenesis, *i.e.*, formation of glucose from noncarbohydrate substances (protein, etc.).

The liver is the organ that forms the blood sugar. This has been proved in many experiments. For example, if the liver is removed completely without disturbing the circulation,[1] the blood sugar immediately commences to fall continuously, even in diabetic animals (Fig. 195), and severe symptoms appear when the hypoglycemia is marked. Intravenous injection of glucose will allow the animal to recover; it can

[1] Mann, F. C., *et al.*, *Am. J. Physiol.*, **65**, 403, 1923: **103**, 45, 1933; *Medicine*, **6**, 419, 1927.

again stand and walk, but sugar is being constantly consumed, so that it is necessary to inject glucose continuously at the rate of 200 mg. per kg. per hr. to maintain a normal blood-sugar level in the hepatectomized or eviscerated dog.[1] Hyperglycemia cannot be provoked in any way in hepatectomized animals except by the injection of glucose; *e.g.*, anesthesia, asphyxia, pancreatectomy, the injection of adrenaline or anterior hypophyseal extracts, etc., do not cause the blood sugar to rise.

The liver plays its part in blood-sugar regulation by keeping a balance between the amount of glucose it secretes into the blood and the amount it takes from the blood and stores as glycogen. During digestive absorption, blood in the portal veins has a higher glucose concentration than blood in the suprahepatic veins; therefore the liver glycogen increases. During fasting the opposite conditions are observed (Claude Bernard). Quantitative determinations of the intake and output of sugar by the liver have shown that there is normally a continuous output of glucose. When the blood sugar rises the hepatic secretion of sugar ceases and glucose is retained by the liver (Fig. 196). On the contrary, in hypoglycemia the liver discharges glucose into the blood, thus raising the blood-sugar level. Soskin[1] has made a careful study of this aspect of hepatic homeostasis.

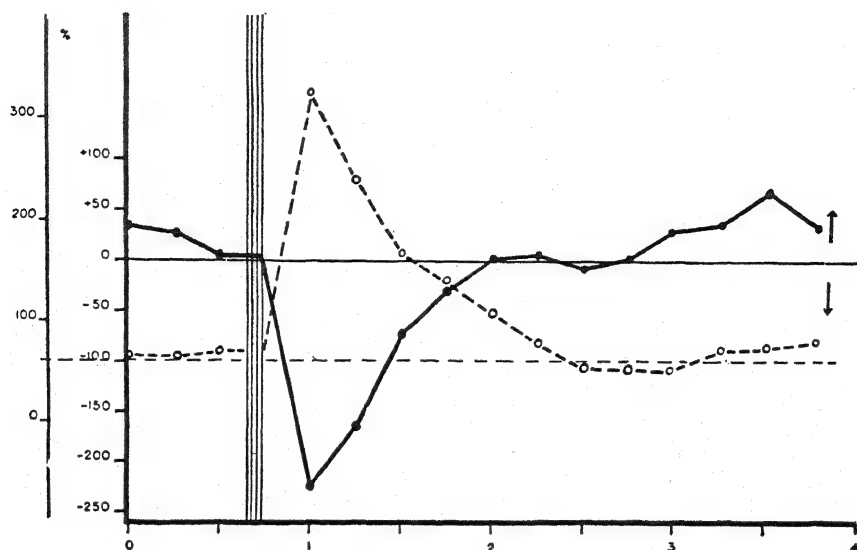



Fig. 196. Hepatic homeostasis. Effect of intravenous injection of glucose (175 gm. per kg.) on the intake and output of glucose by the liver of a dog. Right ordinate, arterial blood sugar, mg. per cent. Left ordinate, liver intake and output, mg. per min. Abscissa, hours. Broken line, arterial blood sugar; heavy solid line, intake (below the 0 level) and output (above) in milligrams per minute. Glucose output ceases immediately when sugar is injected, then considerable amounts are taken up. Throughout the second hour the liver neither retains nor excretes glucose; during this period the blood sugar falls below the initial level and does not rise until the liver again excretes glucose into the blood. Inhibition of hepatic sugar excretion is therefore a real phenomenon, distinct from storage of sugar. (*Soskin, S., et al., Am. J. Physiol., vol. 124, p. 558, 1938.*)

In certain abnormal conditions hepatic homeostasis is disturbed. Thus, in diabetes the liver does not cease to secrete glucose when there is a high blood-sugar level, and sustained hyperglycemia results. Conversely, in hypophyseal or adrenal insufficiency the liver does not increase the secretion of glucose when the blood sugar falls, and severe hypoglycemia is easily provoked by fasting.

A normal blood-sugar level is maintained in spite of the continuous consumption of glucose by the tissues. The amount of glucose stored as glycogen in the liver is enough for only a few hours; therefore blood glucose cannot arise exclusively from the glycogen stored at a given

[1] Houssay, B. A., C. Dosne, and V. G. Foglia, *Am. J. Physiol.*, **141**, 1, 1944.

[1] Soskin, S., *Physiol. Rev.*, **21**, 140, 1941.

moment but must be formed from other non-carbohydrate sources.

Blood in the renal vein sometimes has a higher glucose concentration than arterial blood. This has been considered as proof that the kidney secretes glucose.[1] Further evidence is given by the fact the blood sugar falls more slowly in eviscerated animals if the kidneys are not removed.

Formation, breakdown, and significance of liver glycogen. Hepatic glycogen is continuously being synthesized and broken down, so the glycogen content of the liver varies considerably. The ingestion of carbohydrates increases liver glycogen, and fasting decreases it.

The formation of glycogen (glycogenesis) can be studied in animals in which the store of glycogen in the liver has been previously depleted and reduced to a very low level by fasting, cold, convulsions, phlorhizin, etc. The substance to be tested is then administered to the animal, and the liver-glycogen concentration is again determined.[2]

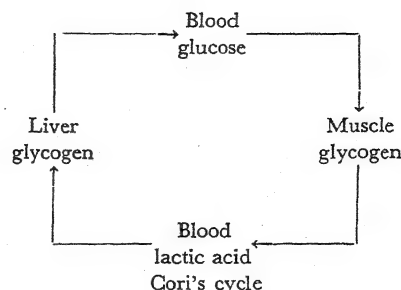
Administration of carbohydrate increases the glycogen content of the liver. Some carbohydrates produce this effect when they are given by mouth or injected into the blood stream (glucose, fructose, mannose, galactose, maltose, pyruvic and lactic acids, glycerol). Others must first be broken down by digestive processes (sucrose, lactose, starch); given by intravenous injection, they cannot be utilized by the organism and are eliminated in the urine. More glycogen is formed from fructose than from an equal amount of glucose; galactose is a much less efficient glycogen precursor.

In some cases an excess of muscle glycogen is broken down to lactic acid, part of which passes into the blood stream and is carried to the liver, where it is converted into glycogen. Hepatic glycogen is then broken down to glucose, which passes into the blood stream and is carried to the muscles, where it is utilized. This process, called "Cori's cycle," is observed in rats and young rabbits injected with adrenaline. It is also observed, although it is not important, during intense muscular exercise.

[1] REINECKE, R. M., *et al.*, *Am. J. Physiol.*, **140**, 276, 1943; 151, 198, 1947.

[2] An increase in liver glycogen does not necessarily signify that a substance administered to the animal has been converted into glycogen, because the increase may have been due to inhibition of the breakdown of glycogen formed from the blood sugar.

Blood glucose → Muscle glycogen → Blood lactic acid → Liver glycogen → Blood glucose

Cori's cycle

Liver glycogen is also formed by the administration of protein and some amino acids, but the rate of synthesis and the amount formed are less than when carbohydrate is given. Glycogen so formed is not originated mainly by direct conversion of protein into glucose or glycogen, because after the administration of glycine with a C isotope to a fasting mouse, the liver glycogen does not contain much C isotope.[1] Also the administration of alanine, marked by an isotope, to a diabetic animal causes the appearance in the urine of an excess of glucose which has little of the isotope in it.[2] Fats give rise to very small amounts of glycogen, which seems to be formed mainly from glycerol, but fatty acids, as has been shown by labeling them with an isotope, can also be converted into glycogen.

The amount of glycogen stored in the liver at a given moment is the balance resulting from the glycogen formed and that broken down. Glycogenolysis (glycogen breakdown) is due to a process of phosphorylation; it ends in the formation of glucose which passes into the blood.

Glycogenolysis occurs in the following circumstances: (*a*) increased consumption of glucose by the tissues (intense and prolonged exercise, cold); (*b*) hypoglycemia (phlorhizin, insulin, fasting, hypophyseal and adrenal insufficiencies); (*c*) acidosis; (*d*) pancreatic diabetes; (*e*) excess thyroid function or administration of thyroxine; (*f*) stimulation of the sympathicoadrenal system; (*g*) asphyxia or anoxia.

A normal concentration of glycogen in the liver is usually a sign of normal metabolism. A decrease in glycogen content renders the liver more susceptible to injury by toxic substances (chloroform, phosphorus); therefore prolonged fasting should be avoided before and after surgical operations. Liver glycogen diminishes

[1] OLSEN, N. S., A. HEMINGWAY, and A. O. NIER, *J. Biol. Chem.*, **148**, 611, 1943.

[2] GURIN, S., and D. W. WILSON, *Federation Proc.*, **1**, 114, 1942.

when the organism does not receive or cannot utilize a sufficient amount of carbohydrate. In this case energy is obtained from fat, and if there is an excessive breakdown of fat, ketone bodies are formed (ketosis and ketonuria of fasting, diabetes, etc.). The metabolism of some amino acids is also disturbed when there is not enough liver glycogen.

Muscle glycogen varies much less than liver glycogen. It decreases a little in hypoglycemia and increases when glucose is injected or there is hyperglycemia; insulin also increases it. Muscle glycogen cannot be utilized to maintain a normal blood sugar after the liver has been removed. Muscle glycogen falls rapidly to a low level during fasting in hypophyseal and adrenal insufficiency (not so rapidly in the latter). Decrease of muscle glycogen can be controlled by administration of the hypophyseal growth hormone; also of certain corticoadrenal hormones (glycocorticoids).

Muscular contraction causes a sharp and marked decrease in muscle glycogen, especially if there is intense, prolonged, and repeated contraction, or a sustained tetanus. During rest muscle glycogen is rapidly resynthesized. In pancreatectomized or in adrenalectomized animals, muscle-glycogen resynthesis during rest proceeds at a much slower rate than in normal animals (Fig. 198). Insulin and some of the corticoadrenal hormones increase the rate of muscle-glycogen formation and sometimes raise the basal level of muscle-glycogen content.[1]

In endocrine insufficiencies (pancreatic, adrenal, etc.) and in hyperthyroidism, disturbances in muscle and liver glycogen appear in the following order: (*a*) liver glycogen decreases, and little or none is formed after glucose injections; (*b*) less muscle glycogen is formed after glucose injection, and the rate of muscle-glycogen resynthesis after it has been depleted by tetanization is decreased; (*c*) in an advanced stage the basal or resting level of muscle glycogen is also decreased.

The Role of the Kidney

There is no appreciable quantity of glucose in normal urine,[2] but when the blood sugar rises to a level near 170 mg. per cent (from 150 to 190 mg. per cent in different cases) glycosuria occurs. The blood-sugar level at which glycosuria occurs is known as the renal threshold for glucose.[1] The following process takes place in the kidney: (*a*) glucose is ultrafiltrated through the glomerulus and is found in glomerular fluid in the same concentration as in blood plasma; (*b*) glucose is then completely reabsorbed by the proximal convoluted renal tube (Richards, Walker). The maximum capacity of tubular reabsorption of glucose (Tm_G) is approximately 375 mg./min. in man and 303 mg./min. in woman (Smith). If the amount filtered by the glomerulus exceeds this maximum, glycosuria occurs. In cases of diabetes that have lasted for some time, the maximum reabsorption capacity may be above the normal and glycosuria does not occur in spite of the existence of moderate hyperglycemia (see pages 726 and 737).

Certain drugs, *e.g.*, phlorhizin,[2] suppress tubular reabsorption of glucose and therefore provoke glycosuria. The effect of phlorhizin is attributed to the inhibition of phosphorylation of glucose in the renal tubes, a process which is necessary for it to be reabsorbed. The role of the kidney therefore consists in preventing the loss of glucose filtered through the glomerulus, thus avoiding a serious metabolic disturbance.

Phlorhizin injections cause intense glycosuria, which persists in the fasting animal. Glucose thus excreted comes from carbohydrate ingested, glycogen broken down, and glyconeogenesis. In prolonged fasting its only source is glyconeogenesis. Excess protein catabolism thus caused is followed by an increase in the excretion of urinary nitrogen. The increase in fat consumption causes a large production of ketonic bodies, which cannot all be metabolized; the excess accumulates in the blood or is excreted in

[1] Endocrine regulation of muscle-glycogen formation has been studied by Foglia, Fernández, Mazzocco, Dambrosi, and Leloir (see *Rev. Soc. argent. de biol.*) and by Long, Lukens, Kendall, Ingle, and others.

[2] The concentration of glucose in normal urine is always below 2 mg. per cent (van Slyke); there are reducing substances in urine which contribute to its slight reducing effect. Fehling's and Benedict's solutions are not appreciably reduced by normal urine.

[1] When there is hyperglycemia and the blood sugar falls, glucose is excreted by the kidneys even when the blood sugar has fallen to levels below 170 mg. per cent; on the other hand when blood sugar rises the threshold is higher and in some abnormal cases it can be as high as 220 mg. per cent. The renal threshold for a rising blood sugar is therefore higher than that for a falling blood sugar.

[2] Phlorhizin is a glucoside extracted from the cortex of apple-tree roots; on hydrolysis it splits into glucose and phloretine.

the urine. This condition is known as "phlorhizin diabetes" because there is glycosuria, ketonuria, and excess urinary N. There are several differences between this type of diabetes and human or other experimental diabetes: (*a*) in phlorhizin diabetes the blood sugar is below normal and there is no hyperglycemia as in other diabetic states; (*b*) administration of sugar is followed by a greater consumption of sugar. In phlorhizin diabetes insulin is secreted by the pancreas, but insulin secretion may be depressed (Houssay and Foglia). When glucose is given to an animal with phlorhizin diabetes, a large part of it is excreted in the urine. The administration of glycogenic amino acids (those which form glucose) is followed by an increase in glycosuria; administration of ketogenic amino acids (those which give ketone bodies) is followed by an increase in ketonuria (see Chap. 43, Protein Metabolism).

In certain human cases of so-called renal diabetes there is glycosuria with normal blood-sugar levels. This disturbance is attributed to deficient reabsorption of glucose by the renal tubes. Some cases of slight glycosuria in pregnancy are apparently produced by a similar mechanism. In pregnant women glycosuria is provoked by small doses of phlorhizin which do not cause glycosuria in non-pregnant women.

ENDOCRINE REGULATION OF CARBOHYDRATE METABOLISM

All living cells can utilize glucose, but in higher animals the production and consumption of glucose, and therefore the blood-sugar level, are controlled by several endocrine factors. These are the secretions of the pancreas, anterior hypophysis, adrenals, and thyroid; other glands play a lesser part. These endocrine glands do not function in an isolated way; they are coordinated in a regulatory mechanism, so that carbohydrate metabolism is controlled by an endocrine balance. Hormones secreted by these glands regulate the storage and secretion of glucose by the liver and the consumption of glucose by the tissues, and they contribute to the maintenance of the normal blood-sugar level and of the normal glycogen concentration in liver and muscle. Reciprocally the blood-sugar level controls the secretion of insulin by the pancreas and that of glucose by the liver, and perhaps the secretions of other endocrine glands that take part in the regulation of carbohydrate metabolism.

The Role of the Pancreas

The importance of the pancreas in carbohydrate metabolism was brilliantly demonstrated by von Mering and Minkowski (1889) when they showed that pancreatectomy in the dog causes a severe diabetes, usually fatal in 1 or 2 weeks, although sometimes in 4 weeks or more. If 12 per cent (one-eighth) of the pancreatic tissue is left in the dog, and 5 to 10 per cent in the rat, the animal does not become diabetic and the blood sugar remains normal. A pancreas thus surgically reduced is not as resistant as the whole pancreas, and it is easily damaged by many agents, such as the administration of anterior hypophyseal extract[1] or thyroid.[2] In some cases—frequently, in the rat—the remaining pancreatic tissue degenerates spontaneously and the animal, after a few months, gradually becomes diabetic; the severity of the condition usually increases as time passes, but sometimes there is spontaneous recovery and the blood sugar returns to a normal level.

The removal of more than seven-eighths of the pancreas in dogs provokes a mild type of diabetes, known as Sandmeyer's diabetes. This condition is not stationary; either the remaining islets degenerate progressively and the severity of diabetes increases until the animal dies, or else the islets improve and increase, the blood sugar eventually returning to normal.

Diabetes following pancreatectomy has been observed in fishes, amphibians, reptiles, birds, and mammals, but there are considerable differences in the severity of the disorder and the length of survival in different species. In the dog and cat this type of diabetes is very severe and rapidly fatal. It progresses more slowly in the rat, monkey, and pig. In some birds (owl) it is very severe; in others (duck) it is so mild that there is scarcely any metabolic disturbance.

Total extirpation of the pancreas has been performed for therapeutic purposes (cancer, etc.) in some twenty patients; severe diabetes occurred in all these cases.

Pancreatic diabetes is due to the elimination of the function of the Langerhans islets.[3] The β cells of the islets secrete insulin, the pancreatic hormone. Damage of the β cells causes diabetes. These statements are based on the following facts:

[1] Houssay, Biasotti, and Rietti, 1932.

[2] Houssay and De Finis, 1943; Houssay, 1944 and 1945.

[3] Diamare, 1889; Laguesse, 1893.

1. Ligature of the pancreatic ducts provokes atrophy of the exocrine (zymogenous) cells, usually without damaging the islets; this operation does not cause diabetes.[1]
2. Repeated injections of anterior hypophyseal extract produces hypophyseal diabetes; if the treatment is sufficiently prolonged, diabetes becomes permanent (metahypophyseal diabetes). In the first case the β cells are severely damaged, and in the second there are irreversible lesions in these cells, which lead to widespread atrophy of the islets. The exocrine cells remain in good condition.
3. Certain toxic substances, such as alloxan,[2] produce lesions in the β cells of the islets, destroying many of them, but leaving the α cells undamaged; these substances produce diabetes.
4. Prolonged hyperglycemia, maintained for several days in cats by intraperitoneal injections of glucose, causes hydropic lesions in the β cells and in some cases permanent diabetes.
5. In all types of experimental diabetes in which the pancreas is not removed, severe and widespread damage to the β cells is observed. Lesions progress by the following stages: degranulation, hydropic degeneration, atrophy, and finally disappearance of the β cells.
6. In 80 per cent of human cases of diabetes, lesions of the β cells have been observed. Lesions are found in these cells in only 16 per cent of nondiabetic subjects. Insulin controls hyperglycemia, glycosuria, and acidosis in all forms of human diabetes.
7. Considerable amounts of insulin have been extracted from the giant islets of certain fishes and from tumors formed by islet tissue in human cases of hyperinsulinism.[1] Some of these tumors are malignant, and metastases of the same tissue in the liver yield large quantities of insulin, which is not obtained from the surrounding hepatic tissue.

The following experiments demonstrate the secretion of insulin by the pancreas: (*a*) pancreatectomy causes diabetes (Fig. 197); (*b*) a pancreatic graft, made by joining the pancreatic artery and vein to the carotid and jugular vein of a diabetic dog, provokes a decrease of the blood sugar to the normal level; (*c*) if the graft is made immediately after pancreatectomy while the blood sugar is normal, hyperglycemia does not occur (Fig. 197); (*d*) this graft also restores the normal concentration of glycogen in the liver and muscle (Fig. 198), secretes adequate amounts of insulin, and performs all the endocrine functions of the pancreas;[2] (*e*) injection of pancreatic venous blood produces a fall in diabetic hyperglycemia, whereas injection of blood from the general circulation does not have

Fig. 197. Effect of a pancreatic graft on the blood sugar. Each curve represents the average blood sugar of several dogs. Left curve, three pancreatectomized diabetic dogs; the arrow indicates the establishment of the graft. Middle curve, four pancreatectomized dogs which were grafted immediately after the removal of the pancreas. Right curve, five pancreatectomized dogs which did not receive a pancreatic graft. (*Houssay, Lewis, and Foglia, Rev. Soc. argent. de biol., vol. 4, p. 859, 1928.*)

[1] Sobolew, 1902.
[2] Shaw, Dunn, *et al.*, 1943.

[1] Excess insulin secretion with symptoms of hypoglycemia.
[2] Delezenne, Guillaumie, and Gayet, 1927 to 1933; Houssay, Lewis, Foglia, Smyth, and De Finis, 1928 to 1942.

this effect;[1] (*f*) daily injections of insulin and an adequate diet have permitted totally pancreatectomized dogs to survive up to 5 years.[2] A normal blood-sugar level has been maintained by insulin treatment in patients who have suffered total pancreatectomy.

Insulin. Insulin is the pancreatic hormone. It was discovered and extracted from pancreatic tissue by Banting and Best in 1922[3] and obtained in crystallized form by Abel (1926). It is a protein; its molecular weight has been reported to vary between 6,000 and 48,000,[4] being made up of units weighing 6,000 or 12,000 each. Determinations of molecular weight of insulin in the solid state made by the x-ray method give a maximum weight of 36,000.[5] This figure agrees with the value of 35,000 to 36,000 found by ultracentrifugation and diffusion of pure insulin solutions.[6] Eleven amino acids have been found in it. It contains sulfur (3.3 per cent), and it loses its activity if the —S—S— bond is broken. It crystallizes as a zinc salt (Scott) and also with cobalt and cadmium. So far it has not been possible to extract insulin from tissues other than the pancreas. International standard insulin has 23 units per mg., but purified insulin containing 27 to 30 units per mg. has been obtained.

Insulin preparations are standardized by biological methods based on their hypoglycemic activity. Two criteria are used for this purpose: (*a*) the average maximum fall in blood sugar produced by the injection of the extract in a series of rabbits; (*b*) the percentage of white mice that have hypoglycemic convulsions after injection of the extract. Sensitiveness of the animals varies, so it is necessary to compare the effect of the unknown extract with that of the international standard insulin.

Insulin secreted by the pancreas has the following physiologic effects: (*a*) it maintains the normal blood-sugar level and prevents hyperglycemia (Fig. 197); (*b*) it controls the secretion of glucose by the liver; (*c*) it regulates the production of glucose from noncarbohydrate sources such as amino acids (glyconeogenesis); (*d*) it maintains a normal liver- and muscle-glycogen concentration (Fig. 198); (*e*) it increases the rate of sugar consumption of the tissues; (*f*) it controls the production of ketonic bodies; (*g*) in diabetic patients it controls most of the specific metabolic disturbances.

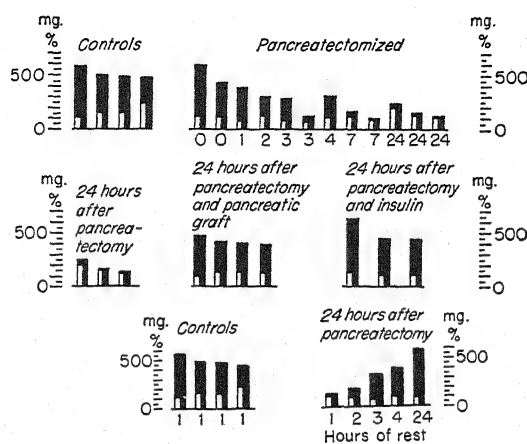



FIG. 198. Resynthesis of muscle glycogen after fatigue. Each column represents the average muscle glycogen of one dog in milligrams per 100 gm. Inset in white, glycogen immediately after fatigue; in black, glycogen after 1 hr. rest, except in the last group where the duration of rest in hours is given under each column. (*Dambrosi, R. G., Rev. Soc. argent. de biol., vol. 9, p. 430, 1933.*)

Insulin is destroyed by the digestive secretions; therefore it loses almost all its activity when given by mouth and must be given by intravenous or subcutaneous injection. These injections cause the blood sugar to fall in normal and in diabetic subjects. The fall is more rapid after intravenous injections and lasts longer after subcutaneous ones. Diabetic patients are usually treated by subcutaneous injections, because they are simpler and have a more prolonged effect. After a few hours the blood sugar rises gradually to the initial level. The level of hypoglycemia and its duration depend on the amount of glucose available, obtained by absorption from the intestine, mobilization of stored glycogen, or production in the liver. There will be little or no hypoglycemia if there are large quantities of glucose available; on the contrary, hypoglycemia will be marked and prolonged if there is little glucose available.

Intense and prolonged hypoglycemia is

[1] Hédon; Zunz and La Barre.

[2] Hédon, MacLeod, Fisher, Bliss, Chaikoff.

[3] BANTING, F. G., and C. H. BEST, *J. Lab. & Clin. Med.*, **7**, 251 and 464, 1922; BANTING, F. G., C. H. BEST, and J. MACLEOD, *Am. J. Physiol.*, **59**, 479, 1922.

[4] GUTFREUND, H., *Biochem. J.*, **42**, 156 and 544, 1949.

[5] HODGKIN, C. D., *Cold Spring Harbor Symposia on Quantitative Biol.*, **14**, 65, 1950.

[6] CREETH, J. M., *Nature*, **170**, 210, 1952.

typical of sensitiveness to insulin; in cases of insulin resistance there is little or no fall in blood sugar. Insulin causes a more marked decrease in blood sugar in subjects previously fed a diet rich in carbohydrates, and a less marked decrease in those fed a diet rich in fat. An intense and prolonged fall in the blood-sugar level gives rise to hypoglycemic symptoms (see Hypoglycemia, page 419).

Insulin increases carbohydrate utilization by the organism. Glucose consumption increases owing to (*a*) increased combustion of glucose (the RQ rises); (*b*) increased formation of muscle glycogen, especially marked if the glycogen store had been previously depleted;[1] (*c*) conversion of glucose into fat. This last is an important effect of insulin, which does not take place in diabetic subjects and is the cause of the scarcity and even complete absence of fat deposit in severe diabetes (Stetten). Plasma phosphate and potassium diminish during insulin hypoglycemia because they are taken up by the tissues. The passage of phosphate to the muscles is particularly marked.[2]

Liver glycogen usually diminishes during insulin hypoglycemia. It increases only in diabetic animals. The decrease in liver glycogen is due to its breakdown caused by the fall in the blood-sugar level. Glucose injections prevent the decrease in liver glycogen, but in nondiabetics there is no more glycogen formed after glucose injections given with insulin than when glucose without insulin is injected.

This mobilization of liver glycogen is due to the direct effect of hypoglycemia on the liver and to the stimulation of the sympathicoadrenal mechanism. The sympathetic nerve centers are stimulated by hypoglycemia and discharge impulses along the hepatic nerves and the nerves to the adrenals. This causes the breakdown of glycogen in two ways—directly, by a nervous effect on the liver, and indirectly, by the discharge of adrenaline from the adrenal medulla. Suppression of the sympathicoadrenal mechanism retards the recovery of the normal blood-sugar level after insulin hypoglycemia.

Insulin diminishes the secretion of glucose by the liver when the blood sugar is normal and also in hypoglycemia. Normal hepatic homeostasis is disturbed by insulin; thus hypoglycemia normally causes an increase in the secretion of glucose by the liver (Soskin), but insulin hypoglycemia does not. Insulin hypoglycemia is not due mainly to an increase in glucose consumption, because intense muscular exercise causes a much greater increase without provoking hypoglycemia. In this case the liver increases the secretion of glucose into the blood, but this does not occur in insulin hypoglycemia. In the diabetic liver, insulin also diminishes the formation of glucose from protein and fat, *i.e.*, glyconeogenesis.

Insulin does not act exclusively on the liver; it also has a direct effect on the tissues. In the eviscerated dog, insulin accelerates the fall in blood sugar; it increases the rate of glucose consumption and the storage of muscle glycogen. This has also been observed *in vitro* in slices of rat diaphragm surviving in Ringer's fluid with added glucose (Gemmill).[1] Muscle can form glycogen when there is no insulin, but at a slower rate than the normal.

Insulin treatment produces a resting condition in the islets, thus favoring regeneration of damaged cells, and protects them from the unfavorable effects of hyperglycemia. The damaged cells can recover if the lesions are not too severe. The histologic aspect, insulin content, and secretion then return to normal. If the damage is so severe that cells are destroyed and the islets atrophied, insulin is no longer efficacious.

The effect of insulin is antagonized by several factors: infections; acidosis; anterior hypophyseal, corticoadrenal, and thyroid hormones; certain anesthetics (anoxia); the existence of anti-insulin; and functional disturbances in the liver. In all these cases there is insulin resistance, and large doses must be given to obtain the usual effects. Probably all these factors modify the sensitiveness of enzymatic systems in the cells. A diet rich in fats requires more insulin than a diet rich in carbohydrate.

The intimate process by means of which insulin exerts its activity is not yet completely understood. Apparently it promotes the passage of glucose from interstitial fluids into the

[1] In order to observe the effect of insulin on muscle-glycogen resynthesis, it is necessary to denervate the muscle, because hypoglycemia may cause convulsions, and muscular contractions thus provoked deplete the muscle-glycogen store.

[2] Sacks, J., *Am. J. Physiol.*, **143**, 157, 1945.

[1] Stadie, W. C., and J. A. Zapp, *J. Biol. Chem.*, **170**, 55, 1947.

cells, with the formation of glucose-6-phosphate catalyzed by hexokinase. It also favors the synthesis of fatty acids through the tricarboxylic cycle, decreasing the oxidation of fats.

The insulin content of the pancreas. There is positive correlation between the number of islets in the pancreas, the condition of the β cells, and the insulin content.[1] The insulin content of the pancreas is diminished during fasting, with a fat diet, or by repeated injections of insulin, but again increases when the animals are fed carbohydrate or mixed diets. Pancreatic tissue remaining after partial pancreatectomy has a normal insulin content; but the insulin content diminishes if the animals are treated with substances that damage the islets, such as anterior hypophyseal extract, alloxan, excess sugar, thyroid, etc.[2] The normal pancreas has approximately 2 units of insulin per gram; the content of normal islets has been estimated at 150 units per gm. The insulin content of the pancreas of diabetic subjects varies considerably; an average of 0.4 units per gram has been reported. In a case of an islet adenoma with symptoms of hyperinsulinism, 85 units of insulin per gram was found. Estrogens, thyroid, and anterior hypophyseal extract in certain doses provoke hypertrophy and hyperplasia of the islets, especially in the rat. It should be noted that large doses of thyroid and anterior hypophyseal extract damage the β cells, causing degeneration and atrophy of the islets, and thus produce diabetes. All these factors—partial pancreatectomy, excess sugar, anterior hypophysis, and thyroid—are supposed to first stimulate the islet cells and then exhaust them, causing degeneration and a decrease in insulin production. Persistent hyperglycemia, rapid growth, a high metabolic rate, and a high caloric diet are factors that concur in exhausting the islets.

Insulin secretion. The concentration of a hormone in a gland does not always give an accurate idea of the amount secreted into the blood. To measure the rate of secretion it is necessary to estimate the concentration in the blood leaving the organ, together with the circulation rate. Insulin determinations in arterial or venous blood have not given completely satisfactory results: they are usually made in diabetic hypophysoadrenalectomized mice or rats, which are known to be very sensitive to insulin. According to Gellhorn[1] the concentration of insulin in the blood of normal subjects is from 0.0001 to 0.0002 unit per cubic centimeter. Insulin disappears from the blood after pancreatectomy in dogs.

The amount of insulin secreted by the pancreas has been estimated by the method of substitution in pancreatectomized animals and in diabetic patients. Continuous intravenous injection of insulin during 12 hr. in recently pancreatectomized dogs maintains a normal blood sugar when the dose is from 0.005 to 0.035 (average 0.017) unit per kilogram per hour; an excessive amount provokes hypoglycemia, and an insufficient amount does not prevent the rise in blood sugar.[2] These doses correspond to a secretion of 30 to 50 units daily for a fasting man of 70 kg. Recent determinations made in pancreatectomized patients have shown that 26 to 80 (average 42) units is required daily to maintain a normal blood sugar when the patient is given 150 to 300 gm. of glucose.[3] It is remarkable that insulin requirements are less after total than after partial pancreatectomy or in many cases of diabetes. The reason for this is not known. Many samples of insulin contain a hyperglycemic substance which can be separated, and there is contradictory evidence that suggests that this substance may be secreted as a hormone by the α cells.

A pancreas grafted into the neck of a dog has been used to explore the mechanism of insulin secretion and has shown that this secretion is controlled mainly by humoral factors and to a lesser degree by nerve impulses. Insulin is secreted continuously, as shown by the fact the blood sugar begins to rise immediately after pancreatectomy (Fig. 197). Insulin secretion is regulated by the blood-sugar level; it increases in hyperglycemia and diminishes in hypoglycemia. Hyperglycemia apparently stimulates the nerve centers and causes insulin secre-

[1] Best, Haist, Campbell, Han, Ridout, Griffith, Marks and Young, Fraenkel-Conrat, Evans, Herring and Simpson.

[2] The insulin content of thyroid-fed animals has not been measured, but lesions of the β cells have been observed, and insulin secretion is diminished.

[1] Gellhorn, E., *et al.*, *Endocrinology*, **29**, 849, 1941; Bornstein, J., and R. D. Lawrence, *Brit. M. J.*, p. 1541, 1951.

[2] Holm, 1927; Houssay, Lewis, and Foglia, 1929; Greeley, 1937 to 1942; Baudoin *et al.*, 1938.

[3] Goldner, M. G., and D. E. Clark., *J. Clin. Endocrinol.*, **4**, 194, 1944.

tion through the vagus nerves;[1] but the main effect is produced by the direct action of glucose on the pancreas. This is shown in the following experiment: Glucose is injected into the artery of a pancreatic graft or of the pancreas *in situ*, and the blood sugar in the general circulation falls, owing to an increase in the secretion of insulin[2] (Fig. 199).

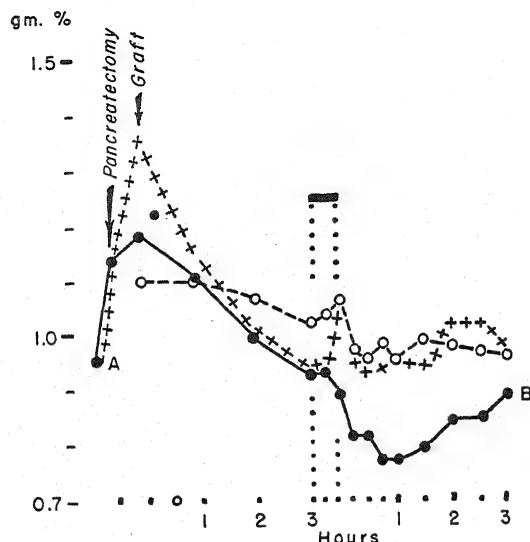



FIG. 199. Direct effect of glucose on insulin secretion. The injection of 2.5 per cent glucose into the artery of a pancreatic graft provoked hypoglycemia (solid line); injection of Locke's solution did not have this effect (crosses); glucose injection in the jugular vein did not produce hypoglycemia (dashes). The horizontal bar represents the time of the injection; the ordinate, blood sugar in grams per cent. (*Foglia, V. G., and R. Fernández, Rev. Soc. argent. de biol., vol. 11, p. 556, 1935.*)

A pancreatic graft maintains a normal blood-sugar level in pancreatectomized dogs. The insulin secretion of the graft is so well controlled that a dog can have three pancreatic grafts beside its own pancreas and still have a normal blood-sugar level. When a pancreas is grafted into a diabetic dog the blood sugar falls to the normal level in 3 to 5 hr., and remains there. To obtain a similar blood-sugar curve by giving insulin to a diabetic animal, it is necessary to inject doses that diminish from 10 to 1 as the blood sugar falls to normal. All these facts demonstrate that insulin secretion controls the blood-sugar level, and reciprocally the blood-sugar level controls insulin secretion.

The extrinsic innervation of the pancreas plays a secondary part in the control of insulin secretion. A pancreatic graft has no nerves, yet it secretes insulin in the varying quantities needed to maintain the blood-sugar level normal and to control hyperglycemia when this occurs. The blood-sugar level is not altered by pancreatic denervation or section of the vagi, nor is the resynthesis of muscle glycogen disturbed. Reports of insulin secretion obtained by stimulating the vagi have been published,[1] but they have not been confirmed by others, in particular by the careful experiments of Etcheverry. Nevertheless the nervous system contributes to the regulation of pancreatic secretion, although it is not indispensable. After pancreatic denervation the return to a normal blood-sugar level following hyperglycemia or hypoglycemia is less rapid than in normal animals (Fig. 200). In the case of hyperglycemia the increase in insulin secretion is delayed; in the case of hypoglycemia inhibition of insulin secretion is delayed.

In all types of diabetes there is a deficiency in insulin secretion. This can be due to (*a*) lesions in the β cells or inhibition of the islets, causing a decrease in insulin secretion (absolute insufficiency); (*b*) resistance of the organism to insulin, although insulin secretion is normal or greater than normal (relative insufficiency).

The capacity of the pancreas to secrete insulin can be established by grafting it into the neck of a diabetic dog. A normal pancreas brings the blood sugar to a normal level in 3 to 5 hr. A pancreas that has had its mass considerably reduced surgically, or that has damaged islets as a result of previous treatment with anterior hypophysis, alloxan, or thyroid, either takes much longer to reestablish a normal blood-sugar level (diminished insulin secretion) or is incapable of controlling hyperglycemia (no insulin secretion).[2]

DIABETES

Symptoms and metabolic disturbances. Total pancreatectomy produces a diabetes of maximum severity. Pancreatic diabetes has a rapid course in the dog, in the cat, and in man,

[1] This fact, reported by Zunz and La Barre, was not confirmed by Gayet.

[2] Grafe and Meythaler, 1927; Gayet, 1928 to 1933; Foglia and Fernández, 1935.

[1] De Corral, Zunz, and La Barre, and others.

[2] Houssay and Foglia, 1936; Houssay, Foglia, Smyth, and De Finis, 1941 to 1943; Houssay, 1944.

but it progresses more slowly in the rat and in other species. The principal signs of diabetes are (*a*) *polyuria*, *i.e.*, an increase in the excretion of urine, a sign which was observed before sugar was discovered in the urine and which is the origin of the name given to the disease; (*b*) *polydipsia*, *i.e.*, increased ingestion of water caused by a strong sensation of thirst; (*c*) *polyphagia*, *i.e.*, increased ingestion of food, caused by a voracious appetite, in spite of which the subject loses weight and gradually becomes emaciated; (*d*) *hyperglycemia*, the blood sugar rising to 200 or 400 mg. per cent; (*e*) *glycosuria*, the urine containing 5 to 10 gm. per cent of glucose. Wounds are easily infected.[1] In severe diabetes there is a high metabolic rate, elimination of nitrogen in the urine increases, fat is rapidly lost, and finally the animals die in cachexia. Before the discovery of insulin the main cause of death of diabetic patients was diabetic coma due to acidosis.

Liver glycogen is considerably diminished. Injection of glucose causes little increase in liver glycogen and in severe diabetes none at all. Muscle glycogen remains normal, but it can diminish in advanced diabetes. The fall in muscle glycogen caused by muscular contraction is similar to that observed in normal animals, but resynthesis of glycogen proceeds at a slower rate.

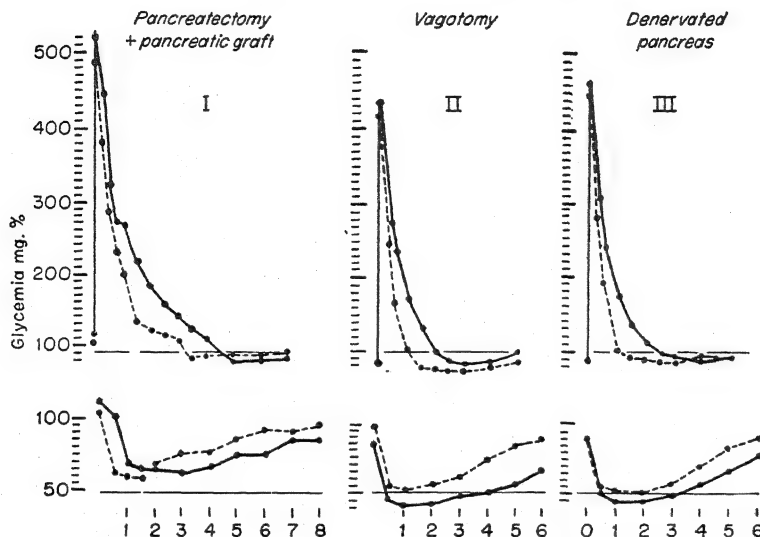



Fig. 200. Effect of pancreatic denervation on the recovery of the normal blood-sugar level. Above: average blood sugar of dogs after intravenous injection of 1 gm. glucose per kilogram. Below: average blood sugar of dogs after intravenous injection of 3 units insulin per kilogram. I: solid line, six pancreatectomized dogs with a pancreatic graft in the neck (chloralose anesthesia); broken line, six chloralosed controls. (*Houssay, Lewis, and Foglia, 1928–1929.*) II: solid line, 12 vagotomized dogs (without anesthesia); broken line, controls (no anesthesia). (*Etcheverry, 1937.*) III: solid line, six dogs with denervated pancreas (no anesthesia); broken line, controls (no anesthesia). (*Etcheverry, 1937.*)

It is remarkable that glycogen increases in the heart muscle and the leukocytes, and sometimes there are glycogen deposits in the renal tubes. The so-called "hydropic degeneration" of the β cells of the islets consists in deposition of glycogen in the protoplasm.

Administration of glucose has the following effects: (*a*) there is a high and prolonged hyperglycemia; (*b*) on the following day the urine contains the same amount of sugar as was injected, or more; (*c*) the RQ does not rise as in normal subjects. These three facts have been interpreted to mean the diabetic organism does not burn sugar. This is not so, because the eviscerated dog burns sugar, a process which, according to Soskin, takes place at a lower rate than the normal at the same blood-sugar levels. Isolated organs, *e.g.*, the heart, consume glucose, although at a lower rate than in normal conditions (Lovatt-Evans). Moreover, after removal of the anterior hypophysis, pancreatectomized

[1] Susceptibility to infection increases with the severity of hyperglycemia and diabetes; nevertheless it is not caused by the excess glucose, but apparently by protein metabolites (Menkin).

animals retain and utilize part of the sugar they receive (Houssay and Biasotti). Hédon has also found an increase in RQ in the final stages of pancreatic diabetes in the dog.

The RQ is approximately 0.7 in diabetics; this has suggested that the diabetic organism burns fat almost exclusively. This interpretation cannot be accepted as final because an RQ of 0.7 may be the result of two simultaneous processes that have opposite effects on the RQ, *i.e.*, oxidation of sugar with RQ = 1 and conversion of fat into sugar with RQ = 0.3.

Diabetic hyperglycemia is dependent on the liver, because it is not possible to provoke any type of diabetes (pancreatic, hypophyseal, etc.) if the liver has been removed. Hepatectomy in a diabetic animal causes the blood sugar to fall rapidly to a low level (Fig. 195), with the appearance of signs of severe hypoglycemia; the animal's condition is relieved by the injection of glucose. Both these facts are further proof that glucose can be utilized in the absence of insulin.

Proteins can give rise to glucose. This process takes place, or at least is begun, in the liver, because this organ is the main site of deamination of amino acids.

In the normal liver, glucose secretion diminishes or ceases when the blood sugar rises, but the diabetic liver has lost this homeostatic property and continues to pour sugar into the blood in spite of hyperglycemia (Soskin). A severely diabetic dog, when fasting, eliminates more sugar than can be accounted for by carbohydrate stored in the body. Urinary nitrogen increases, and the ratio of glucose to nitrogen in the urine (dextrose-nitrogen, D/N, or glucose-nitrogen, G/N, ratio) varies between 2.8 and 3.6. The G/N ratio is taken as an index of the amount of protein converted into glucose. Thus 1 gm. of N is the equivalent of 6.25 gm. of protein; therefore if for every gram of N in the urine there are 3.6 gm. of glucose, 60 per cent of the protein metabolized will have been converted into glucose, which is lost in the urine.[1]

Amino acids have been given to animals with pancreatic or phlorhizin diabetes. The following amino acids provoke the excretion of extra sugar in the urine: glycine, alanine, serine, threonine, valine, glutamic and aspartic acids, ornithine, proline, arginine, cystine, and methionine; they are known as glycogenic amino acids.[1] Other amino acids do not increase urinary glucose, but cause a rise in urinary ketone bodies. The ketogenic amino acids are the following: phenylalanine, tyrosine, leucine, isoleucine, and valine.

An important metabolic disturbance in diabetes is the decrease in the conversion of sugar into fatty acids, catalyzed normally by insulin, to 5 per cent of the normal figure (Stetten).

In severe diabetes fat consumption increases considerably. The greater mobilization of fats causes increased lipemia (to 5 or 10 per cent or even more). Blood cholesterol also rises. The liver has a yellowish color, due to fatty infiltration. Glycerol (10 per cent of fat) is readily converted into glucose. Palmitic acid can also be converted into glucose in the pancreatectomized dog.

In the course of fat metabolism ketone bodies (β-hydroxybutyric acid, acetoacetic acid, and acetone) are produced in the liver and afterward burned in the tissues. When large quantities of fats are metabolized an excess of ketone bodies is produced. Some of them are not burned and they accumulate in the blood (hyperketonemia) and pass out into the urine (ketonuria). In the bladder and in the lung, acetoacetic acid is partially converted into acetone, giving off CO_2, and the breath and urine have the peculiar smell of acetone. β-Hydroxybutyric acid is the most abundant of the ketone bodies. Acetoacetic acid is more toxic, because of its enolic function.[2] Accumulation of these acids causes acidosis. The alkali reserve diminishes, pulmonary ventilation increases, alveolar CO_2 tension falls, urinary ammonia increases, and the pH remains normal (compensated acidosis) or falls (uncompensated acidosis).

Diabetic coma is the outcome of acidosis. At first there is headache, then mental torpor, and finally loss of consciousness. There are nausea and vomiting; abdominal pain and constipation; air hunger, with deep inspirations (Kussmaul breathing); elimination

[1] This ratio is not as constant as it was thought to be. Drury (*J. Clin. Investigation,* **21,** 153, 1942) maintains that the true G/N ratio is from 5 to 7, but that part of the glucose formed is burned; thus the ratio appears to be only 2.8.

[1] According to Bancroft and Drury (*Am. J. Physiol.,* **166,** 213, 1951), if glucose formed and oxidized is taken into account, 90 to 95 per cent of protein can be converted into glucose, and all amino acids would be glycogenic.

[2] Acetoacetic acid has a ketonic $CH_3 \cdot CO \cdot CH_2 \cdot COOH$ and an enolic $CH_3 \cdot COH{=}CH \cdot COOH$ form.

of acetoacetic acid in the urine, which smells of acetone; acidosis (alkali reserve below 25); anhydremia, frequently hypochloremia and increase in nonprotein nitrogen in the blood. There is also a gradual fall in blood pressure. Diabetic coma is treated by the injection of insulin; anhydremia, by intravenous injection of saline solution with or without glucose; and acidosis, by the administration of sodium bicarbonate. Circulatory disturbances should be treated, the stomach should be washed out, and the patient should be kept quiet and warm.

Endocrine pancreatic insufficiency in human diabetes is often slight or moderate in degree, but the severest forms are also observed. In prolonged human diabetes vascular lesions are common. There may be arteriosclerosis with occlusion of the blood vessels, causing gangrene in the limbs, occlusion of the coronary arteries, and lesions in the retina. In some cases the crystalline lens of the eye becomes opaque (diabetic cataract).

Insulin treatment controls nearly all the effects of diabetes. Glucose consumption by the tissues is increased, the excess consumption of fat and protein ceases, liver and muscle glycogen return to normal, hyperlipemia retrogresses, ketosis disappears, and the alkali reserve increases. Each injection of insulin has a marked but only transitory effect, so that repeated doses must be given. A more prolonged effect on the blood sugar, with a more nearly normal blood sugar, can be obtained by injecting insulin preparations that are reabsorbed slowly, such as crystallized insulin, zinc insulin (Scott and Fisher), protamine insulin, protamine-zinc insulin, histone insulin, globin insulin, etc. Protamine insulin[1] is a combination of the two proteins, which flocculate when brought together and then slowly dissociate. This type of insulin has the following advantages: (*a*) the effect is prolonged; (*b*) the glycemic curve is nearer to the normal; (*c*) a smaller number of injections are needed; (*d*) acidosis is more easily controlled; (*e*) hypoglycemic accidents are less frequent. The greater part of insulin used in the treatment of diabetics is protamine insulin, especially NPH 50.

Fortunately insulin is a protein that does not provoke the formation of antibodies, so it continues to produce effects in spite of being injected daily for many years. Resistance to insulin owing to antibody formation has been observed in exceptional cases. Allergic hypersensitiveness to insulin is also exceptional.

[1] Prepared and studied by Hagedorn, Jensen, Krarup, Woodstrup, etc.

Insulin increases activity and prolongs life in diabetic patients. It also permits the survival of patients with juvenile diabetes—a severe, almost certainly fatal, form in the preinsulin era. Death in diabetic coma is now seldom observed.

Hypoglycemia

Causes. The quickest way to produce hypoglycemia is to inject a large dose of insulin. In the rat and rabbit, when the blood sugar falls below 50 mg. per cent, signs of hypoglycemia are observed. Glucose injections are followed by a spectacular recovery from insulin hypoglycemia; within a few minutes all signs of it disappear. Hepatectomy also provokes hypoglycemia, which is equally rapidly controlled by glucose injections. If hypoglycemia is prolonged, the recovery following glucose injections is not so rapid, nor so complete, because of permanent lesions in the nerve centers. These are due to the fact that glucose is the main source of energy of nerve cells.

The causes of hypoglycemia can be classified as follows:

1. Diminished ingestion of sugar: fasting, inanition, cachexia.
2. Deficient absorption of sugar: vomiting, pyloric stenosis, diarrhea.
3. Deficient hepatic homeostasis: hepatectomy, diseases causing extensive damage to the liver, von Gierke's disease.[1]
4. Deficient hepatic homeostasis of endocrine origin: hypophyseal or adrenal insufficiency, occasionally cachexia of hypothyroidism or hyperthyroidism.
5. Hyperinsulinism, *i.e.*, uncontrolled excess insulin secretion.
6. Excess glucose consumption: violent exercise, *e.g.*, marathon races.
7. Excess loss of glucose: renal glycosuria (rare), intense lactation.

Signs. There are many signs and symptoms of hypoglycemia. The following are the most commonly observed: (*a*) flushing of the skin with abundant sweating, or paleness and the sensation of

[1] There is an enlarged liver, with a high glycogen content; glycogen is not readily mobilized and there is hypoglycemia and ketosis. The disease is also called glycogenosis or glycogenic thesaurismosis.

cold; (*b*) palpitations; (*c*) hunger, yawning, epigastric pain, nausea; (*d*) weakness, tiredness, apathy, dizziness; (*e*) anxiety, nervous excitement, aggressiveness, symptoms similar to drunkenness; (*f*) difficulty in performing movements or in speech, tremor, spasms, diplopia; (*g*) convulsions; (*h*) psychic disturbances such as emotional instability, difficulty in concentration, absent-mindedness, disorientation, mental confusion, delirium, amnesia, etc.; (*i*) loss of consciousness, coma. Death may occur if adequate treatment is not given.

Hypoglycemia is treated by giving glucose, fructose, or mannose by mouth. In cases of emergency, glucose should be injected intravenously. Adrenaline and pituitrin are also used, but they are less efficient.

Hypoglycemia and coma (insulin shock) have been provoked deliberately for the treatment of certain mental conditions, *e.g.*, schizophrenia (Sakel's treatment).

Hyperinsulinism is a rare disease in which the pancreas secretes an excess of insulin.[1] Hypoglycemia that occurs on fasting or a few hours after a meal is easily controlled by the ingestion of sugar.

Hypoglycemia is also observed in cases of islet adenoma; extirpation of the adenoma cures the patient.

Insulin treatment has been given with the object of fattening thin persons. The insulin hypoglycemia stimulates the nerve centers of the vagus, thus provoking gastric contractions and secretion, and therefore hunger.[2] Not only does the patient eat more, but the conversion of carbohydrate into fat is also increased.

The Role of the Hypophysis

In acromegaly—a disease due to excess function of the anterior hypophysis—hyperglycemia, glycosuria, and diabetes are frequently found (according to Atkinson in 32.8 per cent of 817 cases collected in the medical literature).[3] This fact had been well known for many years, but the importance of the anterior hypophysis in carbohydrate metabolism has been demonstrated much more recently.

Hypophysectomy. Total hypophysectomy causes many and varied disturbances in carbohydrate metabolism, which are also observed

[1] Harris, 1923; Wilder, 1927.

[2] Bulatao and Cannon.

[3] Atkinson, P. R. B., *Endokrinologie*, **66**, 332, 1938.

after removal of the anterior lobe (pars distalis), but not after that of the posterior lobe.

1. Intestinal absorption of sugar is slowed down,[1] therefore the sugar-tolerance glycemic curve is flattened.
2. After a meal the blood sugar and liver glycogen are normal, but they fall rapidly during fasting[2] and severe hypoglycemia occurs.[3] If treatment by the administration of sugar is not given soon, the subjects die. Hypoglycemia can be prevented by feeding a diet rich in carbohydrate or protein, but not by an exclusively fat diet. Administration of anterior hypophysis or corticoadrenal hormones prevents these accidents. The capacity to maintain a normal blood-sugar level and normal glycogen concentration, to increase muscle glycogen and to control excessive glucose consumption is called by Russell the "glycostatic" function of the anterior hypophysis.
3. Hypophysectomized animals, and those in which the anterior lobe only has been removed, are very sensitive to the hypoglycemic effect of insulin.[4] Small doses of insulin, which have no effect in normal controls, provoke intense hypoglycemia, convulsions, and death. The animals that recover do not regain a normal blood sugar for several hours. There is also increased sensitiveness to other hypoglycemic agents, such as phlorhizin.[5] Loss of hypophyseal growth hormone and adrenal insufficiency contribute to establish this hypersensitiveness. Treatment with anterior hypophysis not only corrects this hypersensitiveness but provokes insulin resistance.[6] Young has called this tne "glycotropic" effect of the anterior hypophysis.
4. Extirpation of the hypophysis, or of its anterior lobe (pars distalis), is followed by marked attenuation of pancreatic and phlorhizin diabetes.[7] Houssay and Biasotti's

[1] Phillips and Robb, 1934. This fact has been amply confirmed.

[2] Houssay and Biasotti, 1930; Russell, Cope, etc.

[3] Wilder in man; Houssay and Biasotti in the dog; this fact has been observed in many species.

[4] Houssay and Magenta, 1924, 1927, and 1929; later amply confirmed.

[5] Houssay and Biasotti, 1930.

[6] Houssay and Potick, 1929; Di Benedetto, 1932; Cope and Marks, 1934; Houssay and Foglia, 1936; Young, 1936 to 1938.

[7] Houssay and Biasotti, 1929, 1930.

Table 42. Blood-sugar Level (Mg. Per Cent) in *Bufo arenarum* (Hensel) in Different Experimental Conditions (Average of Several Hundred Animals)

Condition	*Normal*	*Craneotomy*	*Hypophysectomy*	*Removal of pars distalis*	*Lesion in tuber cinereum*
Pancreas intact...	64	64	50	56	57
Pancreas intact, implantation of pars distalis....	68	69	58	69	60
Pancreatectomy...	199	169*	94	94	127*
Pancreatectomy, implantation of pars distalis....	256	288	228	214	234
Pancreatectomy, implantation of neurointermediate lobe........	199	166	110	116	146

* The slight attenuation of diabetes observed in craneotomized animals, more marked in those with lesions in the tuber, is due to disturbances in the circulation of the hypophysis provoked by these operations, which cause damage and hypofunction of the pars distalis.

strated the "diabetogenic" effect of the anterior hypophysis, its presence increasing the severity of diabetes, while its removal attenuates it. In the toad the diabetogenic effect is produced also after removal of the forebrain or diencephalon, the digestive tract, the lung, the kidneys, or the following endocrine glands: adrenals, thyroid, or gonads. It is not observed after removal of the liver.[1] A diabetogenic activity has been found in the anterior hypophysis of fishes, amphibians, reptiles, birds, and mammals. The anterior hypophysis of man has a particularly strong diabetogenic activity.

In mammals, including man, hypophysectomy also attentuates the severity of pancreatic diabetes (Table 43). The same effect is produced by the removal of the anterior, but not of the posterior, lobe. Comparing hypophysopancreatectomized dogs with pancreatectomized dogs, it is seen that hyperglycemia is not so marked and that the blood sugar rises less after the ingestion of food. There is less glycosuria, less elimination of urinary nitrogen, and the G/N ratio is lower. This decrease in protein break-

Table 43. Effect of Hypophysectomy and Adrenalectomy on Pancreatic Diabetes of Dogs and Cats (Average of Several Animals)

Animal	*Condition*	*Survival, days*	*Urine*			*G/N*	*Glycemia, mg. %*
			Glucose, gm. per kg. per day	*Nitrogen, gm. per kg. per day*	*Ketones, mg. per kg. per day*		
Dog....	Pancreatectomized...........	15	4.0	1.4	60	2.8	380
Dog....	Hypophysopancreatectomized..	74	0.8	1.1	16	0.8	234
		(25–180)*	(0.05–3.2)	(0.4–2.1)	...	(0.7–1.8)	(113–220)
Cat....	Pancreatectomized...........	5	3.2	1.3	116	2.7	347
Cat....	Hypophysopancreatectomized..	22	0.4	0.7	5	0.6	190
Cat....	Adrenalectomized-pancreatectomized	14	0.6	0.6	13	1.0	186

Source: Houssay and Biasotti (dogs); Long and Lukens (cats).
* Figures in parentheses are the extreme values.

(1930) experiments showed that in the toad total hypophysectomy or extirpation of the pars distalis diminished the rise in blood sugar and prevented glycosuria caused by pancreatectomy (Table 42). Implantation of the pars distalis, on the contrary, increased the severity of diabetes to its previous degree or even more. This experiment demon-

down suggests that conversion of protein into sugar is also diminished. Survival is more prolonged, and body weight is lost at a much slower rate. There is less elimination of ketone bodies,[2] probably because of a lesser breakdown of fat. The administration of glucose causes a rise in

[1] Campos, Curutchet, and Lanari, 1933; Foglia, 1942.
[2] Rietti, 1930; Long and Lukens, 1936.

the RQ; glucose is sometimes completely consumed, but usually only in part. It is evident that these animals consume glucose, but at a lower rate than normal animals (Chambers). Slices of surviving tissues taken from these animals consume glucose at a higher rate than the tissues of pancreatectomized animals, but a lower one than the tissues of normal controls. Fasting causes a rapid fall in blood sugar, with signs of severe and even fatal hypoglycemia, which is easily controlled by the administration of glucose.

During the fall in blood sugar, glucose oxidation increases in the hypophysectomized rat (Russell) and rabbit (Greeley), but not in the dog.[1] The fall in blood sugar has been attributed to excessive glucose consumption, but it is due mainly to a disturbance in the function of the liver, which does not increase the secretion of glucose into the blood while the blood sugar is falling, as occurs in normal animals.[2] The relative importance of the two mechanisms is still a moot point. Some workers (Cori, Russell) consider the lack of a moderating action of the hypophysis on glucose consumption to be the most important cause; others consider that the liver does not secrete enough glucose during fasting or when there is hypoglycemia.

The diabetogenic effect of hypophyseal extract. Injection of extracts of the pars distalis of the hypophysis compensates for the effects of hypophysectomy on carbohydrate metabolism, and large doses provoke the appearance of signs opposite to those of hypofunction. This treatment increases the severity of the attenuated pancreatic diabetes in hypophysectomized animals and of diabetes in pancreatectomized animals; also it provokes diabetes in normal mammals.[3] The dose of extract necessary to produce the diabetogenic effect decreases in proportion to the amount of pancreatic tissue extirpated.

Daily intraperitoneal injections of anterior hypophyseal extracts produce in dogs, after 2 to 4 days, a diabetic state (hypophyseal diabetes) with the following characteristics: (*a*) there is a latent period before the appearance of diabetic symptoms; (*b*) there are hyperglycemia, glycosuria, ketonuria, hyperlipemia, and a fatty liver; (*c*) there is marked insulin resistance; (*d*) there is a high liver-glycogen content, except when there is considerable hyperglycemia; (*e*) insulin secretion diminishes 2 to 3 days after the onset of hyperglycemia,[1] and the insulin content of the pancreas also decreases;[2] (*f*) glucose injection is followed by a "diabetic" blood-sugar curve, and there is little or no increase in the RQ;[3] (*g*) hyperglycemia and glycosuria are not observed if the animal is kept fasting or on an exclusively fat diet, but they are observed when the animal is fed a protein diet, and with greater facility if a carbohydrate diet is given; (*h*) the dog is very sensitive to the diabetogenic effect, the rat less, and to obtain it in the toad partial pancreatectomy must be previously performed; (*i*) the smaller the mass of pancreatic tissue remaining in the organism, the easier it is to provoke hypophyseal diabetes; (*j*) hepatectomy causes the blood sugar to fall in animals with hypophyseal diabetes, and this cannot be provoked if the liver has been previously removed.

The growth hormone of the hypophysis has a potent diabetogenic effect and produces transient (Cotes and Young) or permanent (Houssay and Anderson) diabetes. ACTH (adrenocorticotrophic hormone) also has a diabetogenic effect, but it is less potent. None of the other purified hormones has shown this property.

Severe hypophyseal diabetes can be provoked in animals after removal of the hypophysis and the pancreas; therefore it is caused by an extrapancreatic mechanism. In animals with pancreatic tissue the blood sugar rises and there is insulin resistance 2 to 3 days before lesions are observed in the β cells, and insulin secretion diminishes. The damage in the islets is due to (*a*) hyperglycemia; (*b*) a toxic effect of the hypophyseal extract. Hyperglycemia alone is not responsible for these lesions, since they are not observed in animals in which the blood sugar is maintained for several days at the same high level as in hypophyseal diabetes; neither does the effect of insulin diminish in these animals.

During the first few days hypophyseal diabetes can be prevented or cured by insulin treatment (Lukens). If the hypophyseal injections are dis-

[1] Biasotti, Chambers, Soskin, Houssay, Foglia, Dosne; Houssay and Biasotti.

[2] Soskin; Foglia and Potick; Crandal and Cherry.

[3] Evans, Meyer, Simpson, and Reichert, 1932; Baumann and Marine, 1932; Houssay, Biasotti, Di Benedetto, and Rietti, 1932; and others.

[1] Houssay and Foglia, 1936; Houssay, Foglia, Smyth, and Houssay, 1941.

[2] Best, Campbell, and Haist, 1939.

[3] Biasotti, 1934.

continued the diabetic condition disappears in a few days; also if the dose of extract is not gradually increased it may become inefficacious after 8 or 10 days' treatment. If, on the contrary, the hypophyseal diabetic state is maintained for a sufficiently long time, it persists even after hypophyseal treatment is discontinued, and the animal remains with a permanent diabetes. This was first obtained in dogs with partial pancreatectomy (Houssay, Biasotti, and Rietti, 1932) and later in animals with an intact pancreas (Young, 1937). Transitory diabetes observed during hypophyseal treatment is called hypophyseal diabetes (Houssay) or idiohypophyseal diabetes. Permanent diabetes that persists after hypophyseal treatment is discontinued is called metahypophyseal diabetes (Houssay).

Histologic examination of the pancreas made on successive days during anterior hypophyseal treatment shows progressively increasing damage to the β cells.[1] First the cells lose their granulation and there is hydropic degeneration. These lesions are reversible and can be cured spontaneously or by insulin treatment.[2] In this stage insulin secretion and insulin content in the pancreas are diminished. Later the β cells degenerate and disappear; the islets suffer hyaline degeneration and atrophy. At this stage the animal has acquired a permanent diabetes, which is simply a pancreatic diabetes due to insufficiency of β cells in the islets; the exocrine pancreas remains undamaged. In idiohypophyseal diabetes—caused and maintained by anterior hypophyseal extract—there is insulin resistance; metahypophyseal diabetes is due to β cell insufficiency, caused by permanent damage to these cells produced by the injection of anterior hypophyseal extract, which is now no longer necessary to maintain the diabetic state.[3]

Other effects of anterior hypophyseal extract. Anterior hypophyseal extract can stimulate the endocrine pancreas and produce hypertrophy and hyperplasia of the islets,[4] with increased insulin secretion and insulin content in the pancreas.[1] This effect has been attributed to a pancreatotrophic hormone, the existence of which has not been demonstrated. The anterior hypophysis is not necessary for the development and maintenance of the islets; hypophysectomy diminishes growth of the islets (Haist), but it has a more marked effect on the acini, so there is a relative increase in islet tissue, although the total amount is below normal. Insulin content of the pancreas and insulin secretion are not modified by hypophysectomy in the dog.

Certain hypophyseal extracts lower the blood sugar in fasting rabbits and rats, and in man.

The mechanism of the anterior hypophyseal diabetogenic effect. According to Cori and his associates anterior hypophyseal extract inhibits hexokinase *in vitro* and *in vivo;* insulin would antagonize this inhibitory effect. The opposite action of the hormones on the enzyme explains many of the facts, but not all; *e.g.*, it gives no satisfactory explanation of the hypersensitiveness to insulin of hypophysectomized animals. Moreover, purified growth hormone has a potent diabetogenic effect, but does not antagonize hexokinase.

The anterior hypophysis apparently controls carbohydrate consumption and the secretion of glucose by the liver. It stimulates the mobilization and metabolism of fat and the conversion of carbohydrate into fat (see Chap. 42); it increases the formation of protein during growth, and protein breakdown in severe diabetes (see Chap. 43).

The posterior hypophysis. Removal of the posterior lobe does not modify the blood-sugar level, the sensitiveness to insulin, or the course of diabetes. Transitory hyperglycemia is provoked by large doses of posterior hypophyseal extract, which acts on the liver, increasing glucose secretion. Adrenaline increases sensitiveness to this hyperglycemic effect.[2] Posterior hypophyseal extract increases slightly the blood sugar of toads after removal of the pancreas and the hypophysis, but it does not provoke diabetes even in partially pancreatectomized animals.

The Role of the Adrenals

The adrenal cortex. The adrenal cortex plays an important part in carbohydrate, fat, and protein metabolism. Metabolic disturbances observed in adrenal insufficiency are similar to

[1] Richardson and Young, 1938; Richardson, 1940; Porto, 1941.

[2] Lukens and Dohan, 1942.

[3] In some cases of acromegaly there is diabetes with insulin resistance; in others the response to insulin is normal. Probably the former are suffering from hypophyseal diabetes and the latter from metahypophyseal diabetes.

[4] Anselmino and Hoffmann, 1933; Marks and Young, 1939 and 1940.

[1] Houssay, Foglia, Smyth, and Houssay.

[2] Borchardt, 1908; Houssay and Di Benedetto, 1933.

those seen in hypophyseal insufficiency. In adrenalectomized animals and patients with Addison's disease, the blood sugar and liver glycogen fall rapidly after a few hours' fasting. Glucose administration at first restores the normal blood-sugar level and liver glycogen content. As adrenal insufficiency advances, the capacity to form glycogen from injected glucose diminishes first in the liver and later in the muscles; resynthesis of muscle glycogen after fatigue is also retarded (Fig. 198). Previous treatment with corticoadrenal extract prevents the fall in blood sugar and glycogen in fasting adrenalectomized or hypophysectomized animals, and increases the storage of glycogen after glucose administration. Adrenalectomized animals are easily fatigued. Their performance can be improved by the administration of glucose, and much more so by treatment with certain corticoadrenal hormones (Ingle). They fall readily into hypoglycemia when food is withheld or when they are given drugs which provoke hypoglycemia, *e.g.*, insulin,[1] phlorhizin, etc., and when the blood sugar decreases following hyperglycemia provoked by injection of glucose. This hypersensitiveness to insulin is suppressed by factors which in large doses cause insulin resistance in normal and in hypophysectomized dogs: (*a*) adrenocortical extract and glucocorticoids of the adrenal cortex; (*b*) adrenocorticotrophin when the adrenal cortex functions. Desoxycorticosterone and the administration of sodium chloride do not have this effect.

Adrenalectomy diminishes considerably the severity of pancreatic and phlorhizin diabetes, a fact well demonstrated in the cat by Long and Lukens[2] and confirmed in other species[3] (Table 43). Hyperglycemia, glycosuria, ketonuria, urinary nitrogen, and the G/N ratio diminish. Diabetes again becomes severe in adrenalectomized rats when the corticoadrenal tissue is regenerated. In adrenalectomized dogs and cats treatment with corticoadrenal extract or glucocorticoids, but not with sodium chloride, increases the severity of diabetes. Diabetic patients also show a diminished severity in their disease if adrenal insufficiency is established owing to Addison's disease or surgical removal of the adrenals (Thorn, Sprague).

The corticoadrenal hormones with the greatest activity in carbohydrate metabolism have oxygen in an alcoholic or ketonic group in C^{11}. Among the most active are 11-dehydrocorticosterone, 11-dehydro-17-hydroxycorticosterone, and corticosterone; desoxycorticosterone is much less active. These hormones are active in animals and man (Thorn). Large doses (10 mg. per day) produce hyperglycemia and glycosuria in pancreatectomized and normal rats (steroid diabetes, Ingle). They increase liver and muscle glycogen, protein breakdown, and the severity of pancreatic and phlorhizin diabetes which has been attenuated by adrenalectomy or hypophysectomy; also they have an antagonistic action to insulin.

Daily injections of large doses of adrenocorticotrophin in man produce signs of diabetes, such as glycosuria, moderate hyperglycemia, and decreased reabsorption of glucose in the renal tubes. There is also an increase in uric acid excretion and sometimes in the output of nitrogen.[1] So far, permanent diabetes has not been provoked by treatment with corticoadrenal hormones. Nevertheless diabetes has been observed in many patients with adrenocorticoadenoma, and in some cases extirpation of the adrenal tumor has cured the diabetes. A diminished glucose tolerance and transitory diabetes have also been observed in patients suffering from hyperfunction of the adrenal cortex (Cushing's disease, Achard-Thiers syndrome).

Some of the abnormalities in carbohydrate metabolism provoked by adrenalectomy are due to disturbances in sodium, potassium, and water metabolism. The ingestion of NaCl reestablishes the normal rate of glucose absorption which is diminished in adrenalectomized animals. On the other hand, sodium chloride does not prevent hypoglycemia in fasting adrenalectomized animals, nor does it increase the severity of pancreatic diabetes that has been attenuated by adrenalectomy or hypophysectomy. In these cases only corticoadrenal hormones, or substances similar to these hormones, are active.

According to Long the corticoadrenal secre-

[1] This fact has been observed in cats (Cannon), rats (Lewis), and dogs (Lewis and Magenta) and in patients suffering from Addison's disease (Marañon). They are, however, less sensitive than hypophysectomized animals (Bodo, 1952).

[2] Long, C. N. H., and F. Lukens, *J. Exper. Med.*, **63**, 465, 1936.

[3] Lukens in the dog; Hartman and Brownell in the cat; Long in the rat; and Houssay and Biasotti in the toad.

[1] Conn, J. W., *et al.*, *J. Lab. & Clin. Med.*, **33**, 651, 1948: **34**, 255 1949.

tion increases protein breakdown and the formation of sugar from protein. Adrenalectomized animals eliminate less urinary nitrogen in the fasting condition, and less nitrogen and glucose in the diabetic conditions caused by pancreatectomy or phlorhizin. Anoxia increases the elimination of nitrogen and the storage of glycogen in normal fasting rats, but not after adrenalectomy.

Corticoadrenal hormones not only increase glyconeogenesis, but also diminish glucose consumption by the tissues (Ingle) and, probably, the conversion of glucose into fat.

The adrenals are not indispensable for the production of hypophyseal diabetes,[1] but adrenal hormones assist in its development. The anterohypophysis provokes diabetes in two ways: (*a*) by the diabetogenic effect of the growth hormone without the intervention of the adrenals; and (*b*) by stimulation of the adrenal cortex through the adrenocorticotrophic hormone (Long).

The adrenal medulla. Adrenaline injections rapidly provoke hyperglycemia, and muscle- and liver-glycogen breakdown (see page 427, "Role of the nervous system"). While hyperglycemia lasts, glucose is burned in the tissues, but in some experiments a relative decrease in glucose oxidation has been reported. In some animals muscle glycogen is broken down to lactic acid which is reconverted into glycogen in the liver (Cori's cycle). The adrenal medulla secretes continuously a small amount of adrenaline, which does not seem to be of any importance in normal conditions, because removal of the adrenal medulla does not modify the blood-sugar level, nor does it change the course of pancreatic[2] or phlorhizin diabetes. Extirpation of the adrenal medulla, on the other hand, suppresses hyperglycemia produced by a sudden increase in adrenaline secretion (see Chap. 85). Also, recovery of the normal blood-sugar level after insulin is significantly retarded when adrenaline secretion has been suppressed.[3]

The Role of the Thyroid

The internal secretion of the thyroid increases the rate of intestinal absorption of glucose and especially of galactose. In thyroid and anterior hypophyseal insufficiencies absorption of sugar is slow, and in both cases treatment with thyroid or thyroxin improves sugar absorption. In experimental hyperthyroidism sugar is absorbed more quickly than in normal controls. In pa-

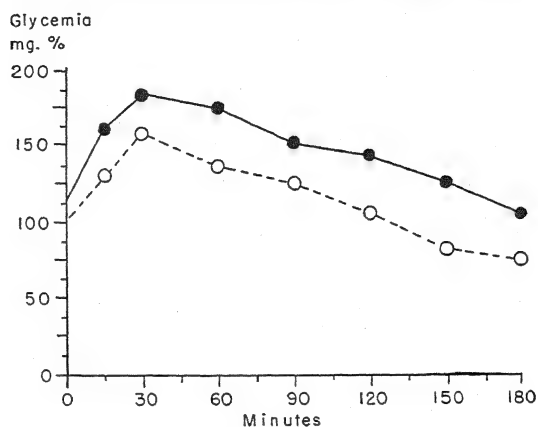



FIG. 201. Blood-sugar curves (average of 11 cases) after ingestion of 1.75 gm. glucose per kilogram while fasting. Solid line, hyperthyroid subjects before operation; average BMR 62 ± 17. Broken line, the same subjects after thyroidectomy; average BMR 3 ± 3.3. Before operation the RQ rose from 0.74 to 0.97, and after operation from 0.78 to 0.99. Before operation 38.4 per cent of the glucose ingested was consumed, and 19.4 per cent after operation. (*Yriart, M., and H. Gotta, Rev. Soc. argent. de biol., vol. 6, p. 1604, 1932.*)

tients with hyperthyroidism increased velocity of absorption has also been observed; after ingestion of glucose (Fig. 201) or galactose (Fig. 202) the blood-sugar curve rises to higher levels than in normal subjects.[1] Intravenous injection of galactose produces the same type of glycemic curve in normal and hyperthyroid subjects. After extensive thyroidectomy—sufficient to control hyperthyroid symptoms—normal postabsorptive glucose and galactose blood-sugar curves are obtained.

In cases of hyperthyroidism the fasting blood sugar is normal, or occasionally a little above normal (120 to 140 mg. per cent); in 15 to 60 per cent slight glycosuria is found, especially after meals. A diagnosis of diabetes should therefore not be made in a hyperthyroid patient unless the fasting blood sugar is above 150 mg. per cent and the blood sugar tolerance curve rises

[1] Houssay and Biasotti (1933–1936–1939); Houssay and Leloir (1935); Houssay, Foglia, and Dosne (1946).
[2] Houssay and Lewis, 1921.
[3] Cannon, McIver, and Bliss, 1923; Houssay, Lewis, and Molinelli, 1924.

[1] ALTHAUSEN, T. L., *J. A. M. A.*, 115, 101, 1940; "Essays in Biology in Honor of Professor H. M. Evans," University of California Press, Berkeley, 1943, p. 13.

above 200 mg. per cent; in nonhyperthyroid subjects a fasting blood sugar of 130 mg. per cent and a sugar-tolerance curve that rises to 180 mg. per cent are considered signs of diabetes.

Sensitiveness to insulin is usually increased in thyroidectomized animals[1] and in many cases of severe myxedema; in hyperthyroidism insulin resistance is frequently increased.

Fig. 202. Galactose-tolerance curves. Average figures of 97 subjects with no thyroid disease (solid line) and of 130 subjects with hyperthyroidism (broken line). (*Althausen, T. L., J. A. M. A., vol. 115, p. 101, 1940.*)

In hyperthyroidism liver glycogen, then heart glycogen, and finally muscle glycogen are considerably decreased. In an advanced stage there is hypoglycemia and hypersensitiveness to insulin. The effect of the thyroid on liver glycogen is antagonized by vitamins A, B, and C, yeast, some fats, and a protein diet.

The endocrine pancreas is stimulated in some animals by thyroid treatment (hypertrophy of the islets and increase in insulin content). Prolonged thyroid treatment, even with large doses, does not seem to cause damage to the normal pancreas; but in dogs with a pancreas reduced by ample surgical extirpation, or by previous treatment with anterior hypophysis, thyroid treatment damages the β cells and the animals become diabetic (thyroid diabetes). If the treatment is not prolonged, the β cells recover and the animal returns to normal, but if it lasts a sufficiently long time the islet lesions become irreversible and permanent diabetes results, which persists even when thyroid administration is discontinued (metathyroid diabetes).[1]

The diabetogenic activity of the thyroid is less intense than that of the hypophysis or the adrenal; nevertheless diabetes occurs twice as frequently in hyperthyroid patients (John) and three times more frequently in patients with adenomatous goiter than in subjects without goiter (Wilder). Diabetes does not provoke hyperthyroidism, which is not more frequent in diabetic subjects than in the general population; but diabetes and hyperthyroidism act reciprocally, each condition increasing the severity of the other. These patients need careful treatment, especially of the hyperthyroid state. Ample thyroidectomy permits the treatment of diabetes as successfully as in nongoitrous subjects. Survival is prolonged; hyperglycemia, glycosuria, and acidosis diminish; and a smaller daily dose of insulin is needed to maintain a normal blood sugar.

Thyroidectomy does not modify the course of pancreatic diabetes in the dog[2] and cat,[3] but it has a marked preventive effect on diabetes provoked by total pancreatectomy in the rat.[4] Thyroidectomy has been performed in diabetic patients who had no thyroid disturbance; carbohydrate tolerance increased and insulin requirement diminished in these patients, but diabetes was not cured. Cases of thyroid insufficiency in diabetics have been reported in which improvement and apparently complete disappearance of the diabetic condition was observed, but the sugar-tolerance curves were still of the diabetic type. Thyroidectomy should not be performed in uncomplicated cases of diabetes to improve carbohydrate metabolism, because it provokes myxedema, which does not add to the patient's comfort, and diabetes can be efficiently treated by other means.

The Role of the Gonads

The incidence of diabetes following removal of 95 per cent of the pancreas in rats is much higher in males than in females. Castration increases incidence of diabetes in subtotally pancreatectomized female rats.[5]

[1] Ducheneau, 1923; Bodansky, 1923; Houssay and Busso, 1924.

[1] Houssay and De Finis, 1943; Houssay, 1944.

[2] Wolfson, 1927; Baronoff, 1928; Yriart, 1930.

[3] Lukens and Dohan.

[4] Houssay and Foglia, 1945.

[5] Foglia, V. G., N. Schuster, and R. R. Rodríguez. *Endocrinology*, **41**, 428, 1947.

Estrogens administered over long periods have a marked protective effect against diabetes in subtotally pancreatectomized male and female castrated rats. Androgens, on the contrary, increase the severity of diabetes and shorten the period of survival after subtotal pancreatectomy.[1] Estrogens cause hyperplasia of the pancreatic islets and inhibit the hypophysis. In rats diabetes has been cured by treatment with estrogens and insulin. Diabetes is more frequently observed in women than in men (Joslin, Marks).

FIG. 203. Regulation of blood-sugar level. Sympathicoadrenohepatic mechanism on the right and vagus-insulin mechanism on the left.

ROLE OF THE NERVOUS SYSTEM

The nervous system assists in the regulation of the blood-sugar level together with the more important humoral factors. Impulses arising in the nerve centers are discharged along two peripheral paths, which modify the blood sugar in opposite directions: (*a*) the sympathicoadrenal-hepatic path, which increases hepatic glycogenolysis and provokes transitory hyperglycemia; (*b*) the vagus-insulin path, which provokes hypoglycemia (Fig. 203). According to the circumstances one of the two mechanisms prevails; usually the sympathicoadrenal is the more active of the two. The suppression of one by section of the corresponding nerves or extirpation of the gland (adrenal or pancreas) leaves the other without its antagonist, and its effects can be more easily demonstrated.

Peripheral stimulation of the splanchnic nerve brings the sympathicoadrenal mechanism into activity. The stimulus spreads along two paths: (*a*) along the hepatic nerves to the liver, where sympathin is set free, provoking glycogenolysis, increased secretion of glucose by the liver, and a rise in blood sugar; (*b*) along the adrenal nerves to the adrenal medulla, provoking secretion of adrenaline, which is carried by the circulation to the liver, where it provokes glycogenolysis, secretion of sugar into the blood, and hyperglycemia. In both cases the sympathetic acts through a chemical mediator: sympathin, which acts locally where it is liberated, or adrenaline, which passes into the general circulation and acts diffusely. Section of the splanchnic nerves suppresses both these paths; section of the liver nerves suppresses the direct (sympathin) effect on the liver, but the adrenal medulla remains active; denervation of the adrenal or extirpation of the adrenal medulla suppresses the effect of adrenaline, but the direct effect (sympathin) remains active (Fig. 203).

Intravenous injection of adrenaline breaks down liver and muscle glycogen. Other drugs provoke this effect indirectly by stimulating adrenal secretion; thus nicotine stimulates the adrenal medulla directly (denervation of the adrenal does not suppress its activity), and adrenaline causes glycogenolysis and hyperglycemia.

[1] LEWIS, J. T., V. G. FOGLIA, and R. R. RODRÍGUEZ, *Endocrinology*, **46**, 11, 1950.

Diffuse stimulation of the nerve centers spreads along both the sympathetic and the vagus nerve paths. Usually the effect of the former predominates and hyperglycemia results. Vagus-insulin activity is made evident by previously cutting both splanchnics or removing the adrenals, in which case hypoglycemia results.

Hyperglycemia due to stimulation of the nerve centers is observed in the following conditions: (*a*) emotion, anesthesia, stimulation of sensory nerves; (*b*) stimulation of the hypothalamus; (*c*) section at the level of the anterior colliculi; (*d*) section of the nerve stem at several levels, principally at the pons in the rabbit (Donhoffer and MacLeod); (*e*) piqûre (stimulation) of the floor of the fourth ventricle, mistakenly called diabetic piqûre by Claude Bernard;[1] (*f*) stimulation of the medulla or the spinal cord.

The effects of stimulation of the vagus have been considered in the paragraphs on insulin secretion.

The nervous, or neurohumoral, mechanisms that regulate the blood sugar are accessory to the more important humoral mechanisms. They are not indispensable, because the blood-sugar level does not change after section of the splanchnic or vagus nerves or after extirpation of the adrenal medulla. These operations do not change the course of experimental diabetes. The nervous mechanisms become active in situations of emergency, such as hypoglycemia, rage, fear, stimulation of the nerve centers, etc. In these circumstances they contribute significantly toward the increase of blood sugar.

Stimulation of efferent nerves does not have a direct effect on the blood sugar, but there is an increase in the glucose consumption of the muscles and glands innervated by the stimulated nerves. In muscular contraction glycogen is broken down and lactic acid produced.

METABOLIC SIGNIFICANCE OF CARBOHYDRATES

Carbohydrate food is the most readily available and cheapest source of energy that can be utilized by man. Approximately 500 gm. can be stored by the human body. This amount covers the need of only one day; therefore it must be given in the form of carbohydrate food or produced from other foodstuffs or body stores. Protein and fat stores are far more important than the meager carbohydrate store which is the equivalent of about 2000 kg.-cal. Carbohydrates "spare" protein, and when they are available less protein is catabolized; this is especially evident in protein fasting, *i.e.*, when there is no protein in the food ingested. Carbohydrates are easily converted into fat. When there is no carbohydrate available, either because the diet is deficient in this type of food or because the body cannot metabolize it (diabetes), more fat is catabolized and large amounts of ketonic bodies are formed (ketosis) which are excreted in part but are also accumulated in the organism, provoking acidosis. Carbohydrate is the principal source of energy mobilized in the course of muscular exercise.

[1] Bernard, C., "Leçons de physiologie expérimentale," J. C. Baillière et fils, Paris, 1855; "Leçons sur la physiologie et la pathologie du système nerveux," J. C. Baillière et fils, Paris, 1858.

BIBLIOGRAPHY

Advances in Carbohydrate Chemistry.

Barron, E. G., The Mechanism of Carbohydrate Metabolism, *Advances in Enzymology*, 3, 149, 1943.

Bouckaert, J. P., and C. de Duve, The Action of Insulin, *Physiol. Rev.*, **24**, 39, 1947.

Cori, C. F., Enzymatic Reactions in Carbohydrate Metabolism, *Harvey Lect.*, **41**, 253, 1946.

Deuel, H. J., and M. G. Morehouse, Interrelation of Carbohydrate and Fat Metabolism, *Advances in Carbohydrate Chemistry*, **2**, 119, 1946.

Dorfmann, A., Pathways of Glycolysis, *Physiol. Rev.*, **23**, 124, 1943.

Geiling, E. M. K., H. Jensen, and G. E. Farrar, Insulin, in "Handbuch der experimentelle Pharmakologie," **5**, 1937, Springer, Berlin, 1937.

Haist, R. E., Factors Affecting the Insulin Content of the Pancreas, *Physiol. Rev.*, **24**, 409, 1944.

Houssay, B. A., Advancement of Knowledge of the Rôle of the Hypophysis in Carbohydrate Metabolism during the Last Twenty-five Years, *Endocrinology*, **30**, 884, 1942.

Houssay, B. A., "Acción de la Tiroides sobre el Metabolismo de los Hidratos de Carbono y en la Diabetes," El Ateneo, Buenos Aires, 1945.

Houssay, B. A., and V. Deulofeu, La Chimie et la sécrétion de l'insuline, *Ergebn. Vitamin- u. Hormonforsch.*, **2**, 298, 1939.

Jensen, H. F., "Insulin, Its Chemistry and Physiology," Oxford, New York, 1938.

Jensen, H. F., Insulin, Its Chemistry and Physiology, *Rev. canad. de biol.*, **5**, 281, 1946.

Krebs, H. A., Oxidation of Carbohydrates, *Advances in Enzymology*, **3**, 191, 1943.

Long, C. N. H., Carbohydrate Metabolism, in Duncan, "Diseases of Metabolism," Saunders, Philadelphia, 1947.

LONG, C. N. H., *Harvey Lect.*, **32**, 194, 1937.

LONG, C. N. H., A Discussion of the Mechanism of Action of Adrenal Cortical Hormones on Carbohydrate and Protein Metabolism, *Endocrinology*, **30**, 870, 1942.

LONG, C. N. H., Adrenals and Carbohydrate Metabolism, *Ann. New York Acad. Sc.*, **43**, 383, 1943.

LUKENS, F. D. W., Alloxan Diabetes, *Physiol. Rev.*, **28**, 304, 1948.

McKEE, F. W., and W. B. HAWKINS, Phlorizin Glycosuria, *Physiol. Rev.*, **25**, 255, 1945.

SOSKIN, S., and R. LEVINE, "Carbohydrate Metabolism," University of Chicago Press, Chicago, 1946.

STETTEN, DE W., The Endocrine Regulations of Carbohydrate Metabolism, *J. A. M. A.*, **132**, 373, 1946.

STETTEN, DE W., The Altered Rates of Certain Metabolic Processes in Hypo- and Hyperinsulinism, *Proc. Am. Diabetes A.*, **7**, 67, 1948.

WATERS, E. T., and C. H. BEST, The Pancreas as an Organ of Internal Secretion, in "Glandular Physiology and Therapy," American Medical Association, Chicago, 1942.

WOOD, H. G., The Fixation of Carbon Dioxide and the Interrelationship of the Tricarboxylic Acid Cycle, *Physiol. Rev.*, **26**, 198, 1946.

YOUNG, F. G., The Anterior Pituitary Gland and Diabetes Mellitus, *New England J. Med.*, **221**, 635, 1939.

YOUNG, F. G., The Pituitary Gland and Carbohydrate Metabolism, *Endocrinology*, **26**, 345, 1940. YOUNG, F. G., *Brit. M. J.*, **2**, 894, 1942.

CHAPTER 42

Fat Metabolism

The body lipids. Fats, or lipids, are found in living organisms and in organic foodstuffs. They are compounds of fatty acids and other substances. The main properties of fats are (*a*) they have fatty acids in their molecule; (*b*) they are insoluble in water and soluble in organic solvents (ether, chloroform, benzene, etc.); (*c*) they are utilized as foodstuffs by living organisms.

Fatty acids. Fatty acids with a carbon chain up to C_4 (butyric acid) are soluble in water. Up to C_{10} they are volatile, and up to C_{11} they are liquid at room temperature. When the chain has C_{12} (lauric acid) or more carbon atoms they are solid at room temperature.

All the fatty acids in animal organisms have an even number of carbons. They form four groups: (*a*) saturated fatty acids ($C_nH_{2n}O_2$), *e.g.*, stearic ($C_{18}H_{36}O_2$), palmitic ($C_{16}H_{32}O_2$), and myristic ($C_{14}H_{28}O_2$); (*b*) monoethylenic with only one double bond, *e.g.*, oleic ($C_{18}H_{34}O_2$); (*c*) diethylenic with two double bonds, *e.g.*, linoleic ($C_{18}H_{32}O_2$); (*d*) triethylenic with three double bonds, *e.g.*, linolenic ($C_{18}H_{30}O_2$). The double bonds invest these acids with a greater chemical reactivity, and they are more easily oxidized and combine with a larger quantity of iodine than saturated fatty acids. The iodine number, or iodine value, is the amount of iodine that can be combined by 100 gm. of fat; it indicates the number of double bonds in a fatty acid.

Natural fats contain many different fatty acids, which vary from one species to another. Thus 12 fatty acids have been obtained from butterfat, including oleic, palmitic, myristic, stearic, and others of a shorter carbon chain.

Lipids can be divided into (*a*) simple lipids; (*b*) compound lipids; (*c*) sterols; (*d*) hydrocarbons.

Simple lipids. These are esters of glycerol, or another alcohol, and fatty acids. They can be divided into

1. Fats and oils, usually called neutral fats or glycerides. These are esters of glycerol and fatty acids, corresponding to the following formula: $CH_2OR—CH{\cdot}OR^{I}—CH_2OR^{II}$ (R, R^I, and R^{II} are fatty-acid radicals).
2. Waxes, which are esters of fatty acids and an alcohol other than glycerol; they are of high molecular weight.

Compound lipids. These are esters of fatty acids and alcohols combined with other substances. They can be divided into

1. Phospholipids, made up of fatty acids, glycerol, phosphoric acid, and a nitrogenous base. Among the monoaminomonophospholipids the most important are (*a*) lecithin, which on hydrolysis yields two fatty acids (oleic, and palmitic or stearic), glycerol, phosphoric acid, and choline; (*b*) cephalin, in which the nitrogenous base is aminoethyl alcohol (colamine). Sphyngomyelin is the prototype of diaminomonophospholipids; on hydrolysis it yields fatty acids (lignoceric, stearic, etc.), phosphoric acid, and two nitrogenous bases, choline and sphyngosine.
2. Cerebrosides or glycolipids, which are made up of fatty acids (phrenosinic or lignoceric), galactose or another sugar, and sphyngosine. Phrenosin and kerasin are the two most important in this group.

Sterols. These substances are not true fats, but alcohols with a cyclic nucleus (cyclopentenoperhydrophenanthrene). They are found free or in esters with fatty acids known as "steroids." They are extracted by organic solvents to-

gether with the fats, and in their functions in the organism they are related to the fats.

Hydrocarbons. Many natural fats contain squalene, carotene, and other lipids of considerable physiologic importance, such as vitamins A and D.

The functions of fats. Fats provide food with a high caloric value; 1 gm. of fat yields 9.3 kg.-cal, twice the energy yield of 1 gm. of carbohydrate or protein (4.1 kg.-cal). The usual diet in adults has 50 to 130 gm. of fat; it is difficult to ingest more than 150 gm. daily, because these large amounts are repulsive and difficult to digest. Nevertheless it is possible to increase the fat content of the diet by carefully choosing and preparing the food (butter is well tolerated) and by acquiring the habit; thus Eskimos eat large quantities of fat. From 15 to 35 per cent of the calories in a normal human diet are provided by fat; the optimum amount is from 80 to 100 gm. per day, and it should not be less than 1 gm. per kg. per day.

When considering the fat in the diet it is necessary to examine (*a*) whether there is an indispensable minimum; (*b*) the advantages of food fat; (*c*) the optimum amount a well-balanced diet should contain.

It is difficult to say whether a definite minimum amount of fat is necessary. Men (Hinhede, Pirquet) and rats (Mendel) have been kept in apparent good health on diets with very little fat. A small amount is needed to carry the fat-soluble vitamins and certain indispensable fatty acids.

Rats fed on a diet completely free from fat can be kept in good condition if adequate amounts of fat-soluble vitamins and certain fatty acids are added. Without these indispensable fatty acids the animals in some cases do not survive long, growth is retarded, and cutaneous lesions appear.[1] All these disturbances are cured by adding to the diet small amounts of linolic and arachidonic ($C_{20}H_{32}O_2$) acids; linolenic, decosahexenoic, and other acids improve the animals but do not cure all the disturbances. Palmitic, oleic, and stearic acids, which are the most abundant fatty acids in a common diet, can be eliminated without causing any harm. Although animals and men have been maintained in good health on an almost fat-free diet, this has been

[1] Burr, G. O., and M. M. Burr, *J. Biol. Chem.*, **82**, 345, 1929; **86**, 587, 1930; **97**, 1, 1932; Hansen, A. E., and G. O. Burr, *J. A. M. A.*, **132**, 855, 1946.

done only in carefully controlled experimental conditions. Optimum growth is obtained in rats when 20 to 40 per cent of the diet consists of fat. The optimum amount for man[1] has been given as 30 per cent of the diet in weight and 50 per cent in calories.

Starling has remarked that fat has many advantages in human nutrition. It offers a high caloric value in concentrated form. It satisfies hunger more than other foodstuffs. It is absorbed slowly, thus spreading the provision of energy over several hours, which is very suitable when hard physical labor must be performed in a cold climate (*e.g.*, in northern lumber camps).

Fats are the most important reserves of energy in the organism, and they can be stored with a much smaller proportion of water than carbohydrate or protein. They form approximately 12 per cent of the total body weight; therefore for each kilogram of body weight there is about 120 gm. fat, equivalent to almost 1100 kg.-cal.—over 72,000 kg.-cal. for a man weighing 67 kg. This reserve is the main source of energy in prolonged fasting; 85 per cent of the total calories are obtained from body fat in hibernating animals and in conditions of inanition.

Fats are protein sparers, but they are not as efficient as carbohydrates in this respect. Nevertheless when they are associated with carbohydrates a greater protein-sparing effect is obtained than with carbohydrate alone.

Fats are the vehicle of the natural fat-soluble vitamins, especially of vitamins A, D, and E. In some cases they reinforce the activity of thiamine (vitamin B_1) and other vitamins.

Subcutaneous fat is poorly vascularized, and little blood circulates through it. It therefore forms an insulating layer that prevents the loss of heat from the body into the environment. This fat layer is particularly important for heat regulation in aquatic mammals and in races living in polar regions (Eskimos).

It also has a plastic effect, rounding the contours of the body, which is of esthetic value when not excessive.

There is no satisfactory proof that isolated muscles consume fat directly in the course of muscular contraction. On the other hand the intact organism apparently can utilize fat to provide energy for muscular exercise, but it must first be converted into substances that can be

[1] Deuel, H. J., *J. Am. Dietet. A.*, **26**, 255, 1950.

used directly by the muscles. At the beginning of exercise there is a high RQ as when carbohydrate is being burned;[1] later, in prolonged exercise, the RQ decreases gradually. The previous diet of the subject has a great influence on the RQ of exercise. Thus a subject on a carbohydrate diet has an RQ of 0.90, of 0.80 on a protein diet, and of 0.72 on a fat diet. The efficiency of muscular exercise is only 22 per cent in subjects on a high-fat diet and 24 per cent on a high-carbohydrate diet (Krogh).

The heart muscle seems to consume fat, of which it contains enough to provide the energy it needs to continue beating for 6 to 7 hr. In certain experiments the oxygen consumption and energy output of the heart cannot be accounted for by the carbohydrate and protein consumed. In these cases it is thought that energy is provided by the oxidation of fat.

Methods of study. There are several methods for the study of fat absorption, transportation, and storage and of the chemical processes in which they take part. The following are the methods most used:

1. A fat containing a fatty acid which does not exist in the organism and is easily detected, *e.g.*, erucic acid, found in colza oil, is given by mouth. A fat thus "labeled" can be followed throughout the body.
2. Fats can also be "labeled" by iodine or a dye which is combined with them.
3. More recently, fats have been labeled by introducing into their molecule an isotope that occurs in nature, such as deuterium, or an artificial radioactive isotope, such as C^{14} or P^{32}.
4. Substances that are thought to be intermediary products of metabolism are administered, and their chemical transformations are followed.[2]
5. Valuable information can be obtained by perfusing an isolated organ with blood or a nutrient fluid to which the substance to be tested is added. The amount destroyed, if any, the products of its metabolism, etc., are then determined by chemical analysis of the fluid or blood flowing from the organ.
6. Slices of surviving tissue are placed in flasks with a nutrient fluid. The substance that is being studied is added and the flasks are shaken continuously to renew the fluid in contact with the tissue. After varying intervals the amount of substance destroyed or its metabolic products are determined by microchemical methods. This technique has the advantage of using simplified chemical systems and of allowing many determinations to be made simultaneously. It has given important results, but these cannot always be considered valid for the whole organism, where many more factors come into play.

[1] Part of the CO_2 eliminated during this period is due to the passage of lactic acid into the blood which sets free CO_2 from bicarbonate; part also is due to hyperventilation and the lowering of the CO_2 tension in alveolar air.

[2] For example, it is thought possible that at a certain stage in the oxidation of a substance either oxalic acid or acetic acid is formed. Oxalic acid is given and it is eliminated quantitatively in the urine; then acetic acid is given and it is oxidized completely. It is evident that in this case acetic acid may be an intermediary product, but that oxalic acid cannot be because it is not metabolized.

INTESTINAL ABSORPTION OF FAT

Digestive processes split neutral fats into glycerol and fatty acids; they split phosphatides into fatty acids, glycerol, phosphoric acid, and nitrogenous bases. Fatty acids are in part converted into soaps by combining with alkali, but part remains in the acid state, because of the acidity of the contents of the upper reaches of the intestine. The products of the digestion of fats are absorbed by the intestinal mucosa and are resynthesized into glycerides and phosphatides, passing into the blood and lymph as such. Fats in maternal plasma pass through the placenta into the fetus.

There is from 4 to 12 per cent fat in the feces, but all this does not come from ingested food. Experiments in fasting animals and in others fed on a fat-free diet have shown that the intestinal mucosa excretes a small amount of fat—approximately 220 mg. per kg. of body weight per day in the dog.[1] Part of the fat found in feces is in bacteria and in cells shed by the intestinal mucosa.

Almost all the fat ingested (more than 95 per cent in man) is digested and absorbed. Fats with a high melting point are less readily absorbed.

Hydrolysis of fats was considered until recently a necessary preliminary to their absorption. Fatty acids set free by hydrolysis of all fat were supposed to be absorbed in the form of soluble soaps (Pflüger). Two circumstances cast doubts on the accuracy of this interpretation: (*a*) the contents of the upper part of the small

[1] SPERRY, W. M., and R. W. ANGEVINE, *J. Biol. Chem.*, **96**, 769, 1932.

intestine are too acid (pH 6.5) for the formation of soaps, and (*b*) the small amount of base available[1] is not sufficient for all the fatty acid which would be released by hydrolysis of all the fat in the food.

Bile plays an important part in the absorption of fats. Bile acids form complex systems with oleic, palmitic, and stearic acids. These systems are soluble in water, are highly diffusible, and pass easily through organic membranes. The effect of bile on fat absorption is due to this property of bile acids, known as hydrotropism (Neuberg). According to Verzár, fats are completely hydrolyzed; the fatty acids thus released combine with the bile acids and penetrate into the intestinal epithelium, where the bile acids are set free and returned to the lumen of the intestine. The fatty acids are converted into phosphatides and then into neutral fats. This phosphorylation is catalyzed by an enzyme which is found in high concentration in the intestinal mucosa.

Frazer[2] maintains that part of the fats are absorbed without previous hydrolysis. Lipolysis is slowed down *in vitro* when 30 per cent of the fatty acid has been released. Partial hydrolysis of fats produces monoglycerides, diglycerides, and fatty acids which together with the bile acids emulsify nonhydrolyzed glycerides. The only efficient system capable of emulsifying fats in the upper part of the intestine, with the formation of droplets of 0.5 μ average diameter, is constituted by bile acids, fatty acids, and monoglycerides. This system permits the absorption of triglyceride, without the need of complete hydrolysis into fatty acids and glycerol. Absorption depends on the diameter of the droplets; thus paraffin is not finely emulsified in the intestine and is not absorbed, but if it is administered emulsified in droplets of 0.5 μ, negatively charged, it seems that it can be absorbed.

Glycerides and long-chain fatty acids apparently penetrate into the fine channels in the outer border of the cells lining the intestine. Monoglycerides and diglycerides are converted into phospholipids or triglycerides; the latter passing into the lymph in the form of droplets. Fatty acids with a short chain and some of the phospholipids behave as water-soluble substances and pass directly into the blood of the portal system and are carried to the liver. There is, however, no final agreement as to these processes.

[1] SCHMIDT, N. K., *Acta physiol. Scandinav.*, **12**, Suppl. 37, 1951.

[2] FRAZER, A. C., *Physiol. Rev.*, **26**, 103, 1946; *Bull. Soc. chim. biol.*, **33**, 961, 1951.

Detergents, such as polyoxyethilene sorbitan mono-oleate (PSM, Tween 80), which emulsify fats have been used in the treatment of diseases where fat absorption is deficient.

The simultaneous presence of the pancreatic juice and the bile is of great importance for the absorption of fats. Pancreatic juice hydrolyzes fats; bile aids this action and facilitates absorption. In the rabbit the bile duct opens into the intestine about 30 cm. above the opening of the pancreatic duct; absorption of fat and the filling of the lacteals begins only at the level where the pancreatic juice is poured into the gut (Bernard).

If the pancreatic juice is absent, 50 to 60 per cent of ingested fat is found in the feces. If the bile only is absent, 30 to 80 per cent is found in the feces, but in this case it is hydrolyzed. If both pancreatic juice and bile are missing, 80 to 90 per cent of the ingested fat is lost in the feces. Steatorrhea (fatty feces) is also observed when there is an excessively rapid passage through the digestive tract or when the intestinal mucosa is damaged, as occurs in certain chronic diarrheas and in celiac disease.

Sinclair[1] maintains that fats are phosphorylated before being absorbed; they are first converted into phospholipids in the intestinal mucosa and then secreted into the blood as neutral fats (glycerides). Phosphorylation of fats has been demonstrated in many experiments: (*a*) after administration of fatty acids easily identified, such as elaidic acid, these are found in phospholipids; (*b*) after ingestion of fatty acids "labeled" by incorporating iodine or deuterium into their molecule, these are also found in phospholipids; (*c*) after ingestion of radioactive phosphorus (which has been used in similar studies) it is found in phospholipids, first in the intestine and liver, later in tissues and fat stores.[2] Glycerophosphates and choline (Honorato) increase intestinal fat absorption.[3]

[1] SINCLAIR, R. G., and C. SMITH, *J. Biol. Chem.*, **82**, 117, 1929; **111**, 260 and 275, 1935; **121**, 361, 1937; *Biol. Symposia*, **5**, 82, 1941.

[2] PERLMAN, I., S. RUBEN, and L. CHAIKOFF, *J. Biol. Chem.*, **122**, 169, 1937.

[3] VERZÁR, F., and L. LASZT, *Biochem. Ztschr.*, **270**, 24 and 35, 1934; **278**, 396, 1935; **288**, 351, 1936.

According to Verzár, fats and some of the sugars (glucose and galactose) must be phosphorylated to be rapidly absorbed. If phosphorylation is inhibited by monoiodoacetic acid or phlorhizin, absorption of these substances is retarded or inhibited even when glycerophosphates are given. Several arguments have been opposed to this theory, *e.g.*, drugs that inhibit fat absorption also damage the intestinal mucosa and retard the absorption of substances that are not phosphorylated. Verzár thought that cortico-adrenal hormones were necessary for phosphorylation, and that deficiencies in absorption observed in adrenalectomized or hypophysectomized animals were due to their absence. Nevertheless these hormones are not indispensable for absorption, as it is enough to control adrenal insufficiency by the administration of sodium chloride to restore a normal rate of absorption. Disturbances in water and salt metabolism seem to be the main cause of abnormalities in absorption in adrenal insufficiency.

Although phospholipids are formed in the intestinal mucosa, many experiments have proved that fats are absorbed mainly as neutral fats, not as phosphatides.[1] These perhaps act as intermediary substances, but not as end products in the resynthesis of fat (Schmidt-Nielsen, Bloom).

In the course of absorption of fats, fatty droplets appear within and between the cells of the intestinal mucosa; these droplets are not there in the fasting animal. Only part of the intracellular content of lipids is revealed by histologic stains, though sudan reveals more than osmic acid. The fat droplets accumulate gradually within the cell to a maximum in the course of 1 hr. or more. Later they diminish, disappearing in about 6 hr.

While fats are being absorbed the intestinal lymphatics become visible because they are filled with a milky fluid; for this reason, they are called lacteals. The milky aspect of the lymph is due to the numerous fine fatty globules that are being secreted into it from the intestine.

Munk found 60 per cent of ingested fat in the lymph collected from a patient with a fistula of the thoracic duct. The results of this observation have been confirmed in many experiments on animals. It is usually said that 60 to 70 per cent of ingested fat is absorbed by the lymphatics and 30 to 40 per cent by the blood capillaries. No accuracy can be attributed to these figures, because part of the fat in the lymph is not absorbed but endogenous fat; moreover, part of the absorbed fat in intestinal lymph can go into the veins without passing through the thoracic duct.

There are no free fatty acids in lymph—only glycerides of fatty acids, arising partly in the body and partly in ingested fat; these are always found as glycerides even when esters of other alcohols are ingested. When palmitic acid labeled with C^{14} is given to unanesthetized rats, 70 to 92 per cent of the amount absorbed is found in thoracic duct lymph, and 96 per cent in the intestinal lymph.[1]

Fatty acids in the blood plasma and erythrocytes increase during the absorption of fat. Cholesterol and cholesterol esters have a tendency to increase, especially in the erythrocytes. Phosphatides also show some increase; they are in part formed from absorbed fatty acids. The blood plasma takes on a milky aspect, due to the numerous very fine fatty globules suspended in it. These particles, known as chylomicrons or lipomicrons or hemoconia, are fine droplets of glycerides surrounded by a protein layer. Their size is 1 μ or less. They are well visualized on a dark field, and can be counted. Lipemia does not always increase during absorption; when it does not, probably fats are retained by the tissues or rapidly stored as they are absorbed.

Fat is transported from the intestine to the tissues in the form of phospholipids, cholesterol esters, and glycerides. Part of the glycerides are transported in lipoproteins.

Lipemia. Chemical methods for blood-fat analysis have been improved sufficiently to give a fairly accurate picture of the blood fat and its changes. The concentration of blood fat in the fasting subject varies considerably in different species[2] and individuals. Even the same individual has widely different figures from one day to another. Table 44 gives the average concentration of the blood fats observed by accurate workers.

Phosphatides and neutral fats are in approximately the same concentration. The greater

[1] Barnes, R. H., E. S. Miller, and G. O. Burr, *J. Biol. Chem.*, **140**, 140, 233, 241, 247, and 773, 1941; Reiser, R., *J. Biol. Chem.*, **143**, 109, 1942; Cardini, C. E., and M. E. Serantes, *Rev. Soc. argent. de biol.*, **20**. 132, 1944.

[1] Bloom, B., J. L. Chaikoff, *et al.*, *J. Biol. Chem.*, **184**, 1, 1950; **189**, 261, 1951.

[2] Boyd, E. M., *J. Biol. Chem.*, **143**, 131, 1942.

part of blood cholesterol is combined in esters; only a small part circulates as free cholesterol. Fats are combined with the plasma protein and form lipoprotein compounds which, in spite of containing 35 to 75 per cent fat, are soluble in water, giving perfectly transparent solutions (Macheboeuf, Cohn).

liver; (*d*) oxidized; (*e*) excreted in milk, in the skin secretion, and in small quantities in the feces. Two kinds of fat are differentiated:

1. Fat that forms part of the tissues. This fat is constant and is not changed by fasting. It has a high iodine number and a large proportion of phospholipids (Terroine's "*élément constant*").[1]
2. Storage fat, which varies considerably with the diet and is diminished by fasting (Terroine's "*élément variable*"). It is mainly made up of glycerides.

Table 44. Lipids in Normal Human Plasma (Mg. Per Cent)

Investigators	*Number of subjects*	*Cholesterol*		*Phospholipids*	*Neutral fats*	*Total lipids*
		Total	*Free*			
Man and Peters, 1933	12	207	..	229	...	659
Boyd*	..	152	46	145	142	530
Van Slyke *et al.*, 1935†	66	232	82	181	225	735
Thannhauser‡	..	150 to 260	40 to 70	150 to 250	0 to 150	

* BOYD, E. M., *J. Biol. Chem.*, **101**, 323, 1933. Out of a total of 353 mg. per cent fatty acids, 146 are neutral fatty acids, 130 form part of phospholipids, and 77 are in esters of cholesterol.

† PAGE, I., E. KIRK, W. H. LEWIS, W. R. THOMPSON, and D. D. VAN SLYKE, *J. Biol. Chem.*, **111**, 613, 1935.

‡ THANNHAUSER, S. J., *New England J. Med.*, **237**, 546, 1947.

There is hyperlipemia (a rise in blood fat) when there is an increase in the fat being transported from one place to another, as in the following circumstances: (*a*) during the post-absorptive period; (*b*) in prolonged fasting or exercise; (*c*) in subjects under the effects of anesthetics or alcohol; (*d*) during pregnancy and lactation, in subjects of both sexes treated with estrogens, and in females treated with gonadotrophins; (*e*) during the laying season in hens and immediately before and during laying in pigeons; (*f*) after repeated injections of anterior hypophyseal extracts (marked hyperlipemia); (*g*) in severe experimental and human diabetes; (*h*) in nephrosis and anemia. Hypercholesterolemia is observed in nearly all these circumstances, being especially marked in pregnancy, severe diabetes, and thyroid insufficiency. Phospholipids increase during pregnancy and diminish after insulin injections.

THE ORIGIN AND FATE OF FAT IN THE ORGANISM

Fats in the organism have their origin in the fat absorbed from the food and in the conversion of carbohydrate and protein into fats.

Absorbed fats are (*a*) deposited as storage fat; (*b*) incorporated into the constituent fats of the tissues; (*c*) converted and deposited in the

Tissue cells sometimes contain not only constituent fats but also variable amounts of storage fats.

STORAGE FATS

Composition. Storage fats vary in different species; thus sheep fat is firmer than beef fat. These differences are due to the dietary and metabolic characteristics of the various species. Fatty acids in the stored fats come mainly from the food fats, but they are modified to a certain extent, their composition approximating that of the animal's fat.

Stored fat does not remain inert in the tissues, as might be supposed from its name. On the contrary, it is being continuously removed for utilization and replaced by new amounts, and there is unceasing metabolic activity in the fat stores. Chemical analysis shows that there is a balance between fat entering and leaving the

[1] TERROINE, E. F., Fat Metabolism, *Ann. Rev. Biochem.*, **5**, 227, 1936.

stores. The chemical mechanisms of fat storage and mobilization are not yet well known.

About half the storage fat in the body is found in the subcutaneous adipose tissue; 10 to 15 per cent in the renal space; 10 to 15 per cent in the mesentery, in the omentum, and under the peritoneum; and 5 per cent between the muscles. The external aspect, microscopic structure, and chemical composition of these fats vary from one part of the body to the other. The melting point of human fat varies from 0.5 to 41°C. in the different stores. It is usually liquid at 37°C. owing to the large proportion of oleic acid it contains. The average composition of human storage fat is given in Table 45.

Table 45. Composition of Human Subcutaneous Fat

Acids	*Cattaneo and Calandra**		*Cramer and Brown†*
	% of fat	*% of total fatty acids*	*% of total fatty acids*
Myristic (tetradecanoic)	1.4	1.5	2.7
Palmitic (hexadecanoic)	19.8	20.8	24.0
Stearic (octadecanoic)	2.0	2.2	8.4
Arachidic (eicosanic)	0.1	0.1	
Myristoleic (tetradecenoic)	0.4	0.4	0.2
Palmitoleic (hexadecenoic)	3.0	3.2	5.0
Oleic (9-octodecenoic)	36.8	38.7	46.9
Linoleic (9,12-octodecadienoic)	23.6	24.8	10.2
Gadoleic (9-eicosenoic)	2.6	2.8	1.5
Eicosadienoic	3.2	3.3	
Arachidonic (eicosatetraenic)	0.4	0.4	1.0
Erucic (docosenoic)	1.7	1.8	
Iodine number	91.7		68.9
Saponification number	192.4		196.5

Note: There is a small amount of squalene (Cattaneo and Calandra) and approximately 1 per cent cholesterol and 0.5 per cent lecithin; 0.1 per cent lauric acid is also reported (Cramer and Brown).

* Calandra, E., and P. Cattaneo, *Rev. Soc. argent. de biol.*, **24**, 275, 1948.

† Cramer, D. L., and J. B. Brown, *J. Biol. Chem.*, **151**, 427, 1943.

In some animals living in cold climates, the iodine number and melting point of fats increase with their depth from the body surface.[1] This is not observed in animal species not commonly exposed to a cold climate.[1]

[1] Henriquez, V., and C. Hansen, *Skandinav. Arch. f. Physiol.*, **11**, 151, 1901.

Fat arising from food-fat. The composition of storage fat can be markedly modified by giving an excess of certain fats in the food. Thus, the subcutaneous fat of a dog fed on linseed oil remains liquid at 0°C.; it becomes firmer and the melting point rises to 50°C. when the dog is fed on mutton fat, which has a large proportion of stearic acid.[2] The administration of colza oil, which has erucic acid, is followed by the appearance of this acid in the fat stores, although it is not a normal constituent of body-fat.[3] Animals have been fed fatty acids "labeled" with iodine, dyes, or deuterium, and these have later been found in their storage fats. Table 46 summarizes some of Anderson and Mendel's experiments on the influence of ingested fat on the composition of storage fat.

In fasting animals the first fats to be mobilized are liquid fats and fatty acids with double bonds (Anderson and Mendel). Hog breeders have observed that submitting the animals to a short fast before slaughtering them renders their fat more firm, especially if they have been fed a diet rich in oil.

Schoenheimer has fed animals with several fatty acids containing C^{14} in their molecule, so as to identify them in the organism. The greater part of those with more than 10 carbon atoms is stored and later utilized. Fatty acids with a shorter chain are either oxidized without delay or else converted into higher fatty acids.[4]

Table 46. Iodine Number of Fat in Diet and Body-fat in Rats Fed a Diet with 60 Per Cent of Total Calories in Fat

Food	*Iodine number*	
	Food-fat	*Body-fat*
Soya oil	132	123
Corn oil	124	114
Linseed oil	108	107
Bacon fat	63	72
Butterfat	36	54

[1] Anderson, W. E., and L. B. Mendel, *J. Biol. Chem.*, **76**, 729, 1928.

[2] Lebedeff, A., *Pflüger's Arch. f. d. ges. Physiol.*, **31**, 11, 1883.

[3] Munk, L., *Arch. f. Anat. u. Physiol.*, **95**, 407, 1884.

[4] Powell, M., *J. Biol. Chem.*, **89**, 43, 1930.

One fatty acid is readily converted into another. Thus stearic acid (C_{18}) is desaturated into oleic acid, or loses two carbon atoms, giving palmitic acid (C_{16}). The chain of palmitic acid can be lengthened into stearic acid or broken down into acids with a shorter chain. Desaturation can take place in the animal organism, but only as far as one double bond; fatty acids with two double bonds come exclusively from food-fat.

Fatty acids are synthesized by successive condensations of two carbon atoms. If acetic acid containing deuterium is given to an animal, the deuterium will be found in the higher fatty acids of its body.[1]

The fat in milk is also modified by the fats in the food, a fact demonstrated by experiments made in lactating cows and bitches. Thus the iodine number of the milk fat rises from 30 to 70 in a cow fed on linseed oil. The taste of milk is due in great part to the fat in the food.

Conversion of carbohydrate into fat. The old popular belief that carbohydrate food is fattening was demonstrated scientifically first by Liebig (1852) and Lawes and Gilbert (1859) and has since then been repeatedly confirmed. A young suckling pig was slaughtered and its body-fat and protein analyzed; another animal from the same litter was fed on a diet with abundant carbohydrate and after a time was slaughtered and analyzed. The increase in body-fat was more than could be accounted for by the protein and fat in the food; therefore it must have been produced by the conversion of carbohydrate (Table 47).

Fat formed from carbohydrate has a high melting point and a low iodine number, *i.e.*, it is made up of saturated fatty acids. Schoenheimer and his associates have shown that palmitic and stearic acid are formed at a much higher rate than nonsaturated fatty acids. According to McHenry *et al.*,[2] certain vitamins in the B complex are necessary for the conversion of carbohydrate into fat. Thiamine is especially important; riboflavin, pyridoxine, and nicotinic acid play a similar but less prominent part in this process.

Experiments in which glucose labeled with deuterium has been given have shown that only 3 per cent of the glucose given is stored as glycogen, while 10 times this amount is converted into fatty acids (Stetten *et al.;* see Chap. 41). The stages in the chemical process of this conversion are not yet known, although possible mechanisms have been discussed. Its RQ is approximately 1.4. There is deficient formation of fat from glucose (Stetten) or acetate (Brady and Gurin) in diabetes. Insulin accelerates the normal conversion of food carbohydrates into fat; also transport of fat from the liver to the tissues (lipokinetic action).

Table 47. Conversion of Carbohydrate into Fat in Pigs

	Pounds
Total fat increase	71.2
Fat in food	12.4
Fat synthesized	58.8
Carbon in synthesized fat	45.3
Protein in food	64
Body-protein increase	6.5
Protein available for fat synthesis	57.5
Carbon in 57.5 lb. protein	27.4

Carbon in synthesized fat obtained from carbohydrate
$45.3 - 27.4 = 17.9$

Source: Lawes and Gilbert, 1859 and 1877.

Conversion of protein into fat. Sixty per cent of the protein intake can be converted into fat; it is therefore logical to suppose that protein can be a source of storage fat. Conversion of protein into fat has been demonstrated by feeding animals on an exclusively protein diet.[1] Firm fat is formed such as arises from carbohydrate. The process is accelerated by pyridoxine, but not by thiamine. Nevertheless not much fat can have its origin in protein, because deamination of amino acids gives rise to fatty acids with a short chain, which are easily oxidized.

Conversion of fat into carbohydrate. Probably fat can be converted into carbohydrate, since the opposite process (conversion of carbohydrate into fat) takes place in the organism, and glycerol can undoubtedly be converted into carbohydrate. Glycogen and blood sugar do not increase after feeding animals with fatty acids, with the exception of propionic acid; neither does fat ingestion increase the sugar in blood or urine of diabetic animals. The RQ corresponding to the conversion of fat into carbohydrate is 0.3. Such a low figure has been registered over a long period in hibernating

[1] Rittenberg, D., and K. Bloch, *J. Biol. Chem.*, **154**, 311, 1944.

[2] McHenry, E. W., and M. L. Cornett, *Vitamins & Hormones*, **2**, 1, 1944.

[1] Longenecker, H. E., *Biol. Symposia*, **5**, 99, 1941; Hoagland, R., and G. G. Snider, *J. Nutrition*, **18**, 435. 1939.

animals only. Over short periods it might be due to temporary retention of CO_2 and not to carbohydrate formation from fat. The passage of C^{14} from fatty acids into glycogen has been observed, and part of C^{14} in palmitic acid given to diabetic dogs has been found in glucose eliminated in the urine.[1] This, however, may be explained by the fact that fats as well as carbohydrates are metabolized in the tricarboxylic cycle (see Chap. 34).

The Role of the Liver in Fat Metabolism

Many observations made in normal and abnormal conditions show that the liver plays an important part in fat metabolism.

Fat content of the liver. Usually the liver contains 3 to 5 per cent fat; one-half to two-thirds of this is made up of phospholipids; the rest is glycerides. According to the circumstances the liver stores variable amounts of fat, phospholipids, or cholesterol. In certain pathologic cases up to 50 per cent of the liver weight can be fat; no other organ ever contains such a large proportion.

Phospholipids and glycerides in the liver have a greater proportion of unsaturated fatty acids than those found in other tissues. This has been attributed to selective storage of these acids or to the liver having the capacity of introducing double bonds into fatty acids, thus facilitating oxidation.[2] The formation of unsaturated fatty acids has been observed in slices of surviving hepatic tissue.[3] In animals fed with saturated fatty acids labeled by deuterium, unsaturated fatty acids containing deuterium are found.[4]

Phospholipids are formed by the liver at a higher rate than in any other organ, except the intestinal mucosa during absorption, and they are rapidly secreted into the blood; thus the liver is the main source of blood phospholipids. Acetate is converted in the liver into neutral fats and phospholipids; fat is usually formed at a faster rate, so possibly phospholipids are not necessary intermediate substances.[5]

[1] Abraham, S., J. L. Chaikoff, and W. Z. Hassid, *J. Biol. Chem.*, **195**, 567, 1952.

[2] Leathes, J. B., and H. S. Raper, "The Fats," Longmans, New York, 1925.

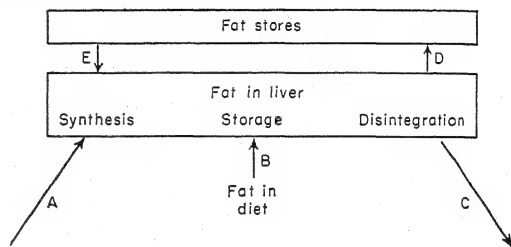
[3] Artom, C., *Ann. Rev. Biochem.*, **3**, 212, 1934.

[4] Schoenheimer, R., and D. Rittenberg, *J. Biol. Chem.*, **113**, 505, 1936; **117**, 485, 1937.

[5] Pihl, A., and K. Bloch, *J. Biol. Chem.*, **183**, 431, 1950.

Ketone bodies are formed in the liver in the course of oxidation of fatty acids (see "Ketogenesis and ketolysis," page 443).

Liver fat and fatty liver. The fat content of the liver depends on the balance between fat formed or stored in the liver and fat secreted from the liver. The fat in the liver has its origin in (*a*) fats in the food (*B* in the diagram); (*b*) other fat stores (*E*); (*c*) synthesis of fat in the liver (*A*). Fat that disappears from the liver is either broken down and oxidized in the hepatic cells (*C*) or secreted into the blood and transported to other fat stores (*D*).

Fat stores
E D
Fat in liver
Synthesis Storage Disintegration
B
Fat in diet
A C

In certain cases the liver is infiltrated by fat. In the so-called fatty degeneration of the liver, intracellular deposition of fat can be observed histologically. Eventually this accumulation of fat disturbs the functions of hepatic cells. A larger amount of fat is revealed by chemical analysis than by observation under the microscope; the difference between the two is called "masked fat."

Among the principal causes of fatty liver the following can be mentioned:

1. Dietary factors: excess of fat, cholesterol, cystine, thiamine, or biotin; prolonged fasting or insufficiency of protein, choline, methionine, or other lipotropic factors.
2. Extensive partial hepatectomy. There is a marked rise in the fat content of the remaining hepatic tissue, which commences a few hours after the operation and lasts several days.
3. Endocrine factors: (*a*) in pancreatic and severe forms of other experimental diabetes, and human diabetes; (*b*) following injection of anterior hypophyseal extract, somatotrophin or adrenocorticotrophin; (*c*) in pregnancy (the glycerides, phospholipids, and cholesterol increase); (*d*) at the beginning of sexual maturation.
4. Toxic factors: phosphorus, chloroform, alcohol, phlorhizin, etc.

5. Infectious or degenerative processes; *e.g.*, in yellow fever.
6. Environment: very high or very low temperatures; hypoxia.

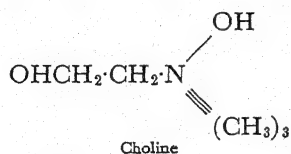
Fatty infiltration creates favorable conditions for degenerative processes and eventually can end in cirrhosis of the liver. On the contrary, carbohydrate protects the liver, but only if there is an adequate amount of protein in the diet. It is usually admitted that fat and carbohydrate accumulation in the liver cells are mutually antagonistic, but there are many exceptions to this rule.

Diets with excess fat and deficient in protein, cystine, or methionine may cause necrosis of the liver, *i.e.*, the death of the hepatic cells.

Lipotropic and antilipotropic factors. Lipotropic factors are substances that prevent the accumulation of excess of fat in the liver, and diminish the fat content of the liver when it is excessive. The principal lipotropic substances are (*a*) choline and substances chemically related to choline, *e.g.*, betaine; (*b*) methionine and proteins containing methionine; (*c*) inositol; (*d*) certain substances extracted from the pancreas.

Antilipotropic factors are substances which oppose the lipotropic effect of choline and other lipotropic factors and which in high doses create conditions favorable to fatty infiltration of the liver. Cystine, thiamine, cholesterol, biotin, etc., are antilipotropic substances.[1]

The lipotropic effect of choline was discovered as follows: Pancreatectomized dogs maintained free from diabetes by insulin treatment could nevertheless not be kept alive for a very long time. A fatty liver was observed in these animals, and the fatty infiltration could be prevented or cured by adding pancreas to the diet. Then it was found that lecithin had this activity,[2] and when the products of its hydrolysis were tested, choline was found to be the active substance (Best *et al.*, 1932 to 1934).

$$OHCH_2{\cdot}CH_2{\cdot}N \begin{cases} OH \\ (CH_3)_3 \end{cases}$$

Choline

[1] McHenry, E. W., and G. Gavin, *J. Biol. Chem.*, **125**, 627, 1938.

[2] Hershey, G. M., and S. Soskin, *Am. J. Physiol.*, **98**, 74, 1931.

Other substances have been shown to have the same effect, among them homocholine (which is even more active than choline), betaine, methionine and proteins containing methionine, etc.[1] All these substances have this activity because they possess methyl groups which are used in the formation of choline. A very small amount of choline (5 mg. in the rat) is sufficient to prevent fatty infiltration of the liver produced by anterior hypophyseal extract or any other cause.

There are lipotropic factors in pancreatic tissues free from choline. An active principle known as lipocaic[2] has been discovered,[3] but it has not yet been obtained in pure form. According to Chaikoff, there is an anti-fatty liver factor (AFL) in pancreatic tissue and pancreatic juice. This factor has proteolytic activity and permits utilization of methionine in proteins.[4]

Dragstedt *et al.* maintain that pancreatectomized dogs suffer from an insufficiency of lipocaic, which can be compensated by the ingestion of pancreas or of lipocaic. This insufficiency has the following results: glycosuria diminishes, in spite of decreased doses of insulin; hypoglycemic crises frequently occur; hepatic functions are depressed; there are low blood fat, general weakness, anorexia, and emaciation; eventually the animal dies. Administration of lipocaic is followed by a decrease in liver fat and an increase in glycosuria and the insulin requirement, while the condition of the animal improves.

There are differences in the chemical mechanisms by which the lipotropic factors produce their effect on the fatty liver, arising from different causes.

Choline and inositol accelerate the formation of phospholipids from liver fat; these phospholipids are then secreted from the liver. Choline is incorporated into the phospholipid molecule. This has been demonstrated by administering choline labeled with arsenic, heavy nitrogen, or radioactive phosphorus, which is later found in lecithin formed in the liver. When there is not enough choline, fat accumulates in the liver because it cannot be converted into phospholipids and thus mobi-

[1] Tucker, H. F., and H. C. Eckstein, *J. Biol. Chem.*, **121**, 479, 1937.

[2] Lipocaic signifies "fat-burning."

[3] Dragstedt, L. R., *et al.*, *Am. J. Physiol.*, **117**, 175, 1936; *Proc. Soc. Exper. Biol. & Med.*, **75**, 785, 1950.

[4] Chaikoff, I. L., and C. Enteman, *Advances in Enzymology*, **8**, 171, 1948; Clowes, G. H. A., *et al.*, *Am. J. Physiol.*, **165**, 628, 1951.

lized. Methionine and proteins containing it produce choline; to this is due their lipotropic activity.

The fatty liver produced by anterior hypophyseal extract is due to the accumulation in the liver of fat coming from other fat stores. The fatty liver caused by insufficient choline is due to the non-mobilization of liver fat. The fatty liver provoked by cystine or thiamine is apparently due to excessive synthesis of fatty acids (Stetten and Salcedo).

Diets deficient in protein can provoke fatty degeneration and eventually cirrhosis of the liver, causing high mortality, especially in children. This nutritional deficiency (kwashiorkor and allied diseases) has been reported in communities in Africa, the West Indies, Chile, etc.

The Role of the Lung in Fat Metabolism

According to Roger and Binet,[1] fat absorbed by the lacteals is carried by the blood to the lungs, where it is fixed (lipopexia) and then disintegrated (lipodieresis). Other workers have found this process to be of secondary importance.[2]

The Role of the Endocrine Glands in Fat Metabolism

Some of the endocrine glands play an important part in fat metabolism, which will be considered in the chapters on the glands in question.

The Role of the Nervous System in Fat Metabolism

Extreme obesity has been observed in rats,[3] dogs,[4] and monkeys[5] after certain experimental lesions were produced in the hypothalamus. This effect is apparently due to hyperphagia; the animals have a huge appetite and eat great quantities of food very rapidly.[6] (See page 1081.)

Transmethylation[7]

It has been proved that in the tissues of several animals, including man, methyl groups can be transferred from one substance to another. A normal diet contains substances that can supply methyl groups for this process. If they are absent, growth ceases, the liver becomes infiltrated by fat, there is hemorrhage into the renal tubes, and eventually death occurs. If a methyl donor such as choline, betaine, or methionine is added to the diet, these disturbances are prevented or cured.

Methyl groups play a part in the metabolism of nitrogen, sulfur, protein, carbohydrate, and fat.[1] Methionine is one of the amino acids essential for the maintenance of a normal nitrogen balance, growth, and even life. Methionine can be formed by transference to homocysteine of a methyl group from choline or betaine. The methyl groups of methionine are used in the formation of choline, which exerts lipotropic activity in the liver. The methyl groups of methionine are also used for methylation of creatine, creatinine, adrenaline, etc. These processes of transmethylation have been demonstrated by labeling the methyl with deuterium or C^{14}. Vitamin B_{12} accelerates methyl biosynthesis.

The Functions of Phospholipids

The liver, heart, kidney, spleen, lung, and other viscera have 2 to 4 per cent ether-soluble extract, of which one-half to two-thirds is phosphatides, the remainder being glycerides and cholesterol. The phospholipid content of a tissue is directly proportional to its functional activity (Bloor). It appears that phospholipids are the main form in which fat is transported in the organism, owing to their solubility in water and high fatty-acid content (70 per cent). They are also highly reactive and easily oxidized substances because they have more unsaturated fatty acids than the glycerides. The liver and the intestinal mucosa are the sites of greatest activity in the synthesis of phospholipids.

The formation of phospholipids is of importance in the absorption of fat, according to some observers, because part of the fat in food is phosphorylated in the intestinal mucosa. Choline and inositol, as has been noted above, prevent the accumulation of fat in the liver because they speed up production and oxidation of phospholipids in the liver cell.[2] Probably most

[1] Roger, G. H., and L. Binet, *Presse méd.*, **26**, 277, 1922.

[2] Hoppe, G., *Ztschr. f. d. ges. exper. Med.*, **89**, 97, 1933; Cantoni, O., *Arch. de physiol. norm. et path.*, **28**, 205, 1930; Lombroso, M., *Rass. clin. ter. sc.*, aff. p. 57, 1935.

[3] Smith, P. E., and C. F. Greenwood, *Anat. Rec.*, **29**, 373, 1925.

[4] Solari, L. A., *Acta V Congr. Nac. Med.*, **3**, 303, 1934.

[5] Ranson, S. W., C. Fisher, and W. R. Ingram, *Endocrinology*, **23**, 175, 1938.

[6] Brobeck, J. R., *Physiol. Rev.*, **26**, 541, 1946.

[7] Du Vigneaud, V., *Proc. Am. Philos. Soc.*, **92**, 127, 1948.

[1] Choline deficiency will be considered in Chap. 49, Vitamins.

[2] Chaikoff, I. L., *Physiol. Rev.*, **22**, 291, 1942; McHenry, E. W., and J. M. Patterson, *Physiol. Rev.*, **24**, 128, 1944.

of the fat entering and leaving the cells is in the form of phospholipids.

It has been remarked that the fat and phospholipid content of certain organs (brain, kidney, lung, spleen, and heart) varies little under the influence of fasting or changes in diet. This has been called the constant or specific constituent fat of each tissue.[1] The term is not quite appropriate because this fat is not static and unchangeable; on the contrary, the constancy of the fat content is due to a dynamic equilibrium between the fat entering and formed in the tissues on one hand, and that leaving and destroyed on the other. The administration of phospholipids with radioactive phosphorus has shown that there is a continuous intake and output of these substances in the tissues. The quickest exchanges are observed in the intestinal mucosa, those in the liver come next, in the heart and muscle they are less rapid, and in the brain they are even slower. The concentration of phospholipids is fairly constant for each organ, but the fatty acids that form part of them vary according to the nature of the fatty acids in the diet.

Apart from these constituent fats (mainly phospholipids), in special circumstances certain tissues, such as muscle, and particularly the liver, can store variable, sometimes considerable, amounts of fat.

Phospholipids are of fundamental importance in the functioning of the cells. They are a permanent constituent of the cell membrane and protoplasm. Their loss causes serious anatomic and physiologic disturbances. The ratio of phospholipids to cholesterol is maintained constant in the cells, and this appears to be important for the permeability of the membrane, hydration of the protoplasm, etc. In some cases phospholipids and cholesterol have opposite effects, and an equilibrium is kept between them in normal conditions. For example, cholesterol neutralizes the intense hemolytic activity of lysocythin and lysocephalin produced by the disintegration of lecithin and cephalin caused by certain phosphatidases, such as that of cobra venom.

Table 48. Oxidation Products of Fatty Acids

Acid ingested	*Oxidation product*	*Acid excreted*
Benzoic ($C_6H_5 \cdot COOH$)	Not oxidized	Hippuric
Phenylacetic ($C_6H_5 \cdot CH_2 \cdot COOH$)	Not oxidized	Phenaceturic
Phenylpropionic ($C_6H_5 \cdot CH_2 \cdot CH_2 \cdot COOH$)	$C_6H_5 \cdot COOH$	Hippuric
Phenylbutyric ($C_6H_5 \cdot CH_2 \cdot CH_2 \cdot CH_2 \cdot COOH$)	$C_6H_5 \cdot CH_2COOH$	Phenaceturic
Phenylvaleric ($C_6H_5 \cdot CH_2 \cdot CH_2 \cdot CH_2 \cdot CH_2 \cdot COOH$)	$C_6H_5 \cdot COOH$	Hippuric

Source: Knoop.

OXIDATION OF FAT AND KETOGENESIS

The final products of fat metabolism are CO_2 and water. Glycerol and fatty acids (the two components of neutral fats) undergo widely different processes. Glycerol follows the same path as carbohydrate, probably after phosphorylation. It can be utilized as a source of glycogen in normal animals, and it increases glycosuria in diabetic animals.

The fatty acids found in greatest amounts in the organism are palmitic acid, which has 16 carbon atoms in its chain ($CH_3 \cdot (CH_2)_{14} \cdot COOH$); stearic acid, with an 18-C chain ($CH_3 \cdot (CH_2)_{16} \cdot COOH$); and oleic acid, with an 18-C chain ($CH_3 \cdot (CH_2)_7 \cdot CH = CH \cdot (CH_2)_7 \cdot COOH$). Oxidation of these acids has been studied mainly by three methods: (*a*) the administration of acids combined with a benzene ring (Knoop's experiments); (*b*) measuring the formation of ketone bodies in the organism or in slices of surviving tissues; (*c*) the administration of fatty acids labeled by an isotope (deuterium or C^{14}).

These methods have shown that fatty acids are oxidized at the β-C (the third C counting from the carboxyl group), and a substance with two carbon atoms is separated. By a series of successive oxidations, the fatty-acid chain thus loses its carbon atoms two by two. Two molecules of the two-carbon substances formed can combine and give ketone bodies that have four

[1] MAYER, A., and G. SCHAEFFER, *J. de physiol. et de path. gén.*, **15**, 510, 1913; **16**, 1, 1914; TERROINE, E. F., "Physiologie des substances grasses," Paris, 1919; *Ann. Rev. Biochem.*, **5**, 227, 1936.

carbon atoms, or they can be oxidized following the tricarboxylic cycle. Ketone bodies are also oxidized by the tissues, and in normal animals only a small amount passes into the urine.

Knoop's experiments. The now classic work of Knoop[1] showed that fatty acids undergo β-oxidation and that the chain loses its carbon atoms by pairs. He gave animals phenyl derivatives with a fatty-acid side chain (Table 48) and observed that, when the fatty-acid chain had an even number of carbon atoms, disintegration of the derivative ended in phenylacetic acid. On the other hand, phenyl derivatives with a side chain of an odd number of carbon atoms ended in benzoic acid. Phenylacetic acid and benzoic acid are eliminated in the urine combined with glycine in the form of phenaceturic and hippuric acids, respectively. The amine group of glycine combines with the carboxyl of phenylacetic or benzoic acid, as in the peptide linkage of proteins. Results obtained by the administration of some phenyl derivatives are summarized in Table 48.

Knoop concluded that the fatty-acid chain was always oxidized in the β-carbon; the chain, breaking at the bond between the α- and β-carbons, is shortened by two carbon atoms at a time. Thus stearic acid, which has 18 carbon atoms, gives palmitic acid with 16 C, which in turn gives myristic acid (14 C), then lauric acid (12 C) is formed, and so on down to butyric acid (4 C), which would be oxidized into ketone bodies, the latter also undergoing final oxidation into CO_2 and water. This theory gives a simple explanation of how ketone bodies are formed, as can be seen in the following diagram:

$$\overset{\omega}{CH_3}\cdot CH_2\cdot CH_2\cdot \overset{\beta}{CH_2}\cdot \overset{\alpha}{CH_2}\cdot COOH \quad \text{Caproic acid}$$
$$\downarrow \beta \text{ oxidation}$$
$$CH_3\cdot CH_2\cdot CH_2\cdot CO\cdot CH_2\cdot COOH$$
$$\downarrow \beta \text{ oxidation and disintegration}$$
$$CH_3\cdot CH_2\cdot CH_2\cdot COOH + CH_3\cdot COOH \quad \text{Butyric + acetic acids}$$
$$\downarrow \beta \text{ oxidation}$$
$$CH_3\ CO\cdot CH_2\cdot COOH \quad \text{Acetoacetic acid}$$

Later work has shown that β-oxidation does take place, but that the final stages are not those postulated by Knoop.

The formation of ketone bodies. Oxidation of fat and of some amino acids in the organism gives rise to the so-called "ketone bodies," *i.e.*, acetoacetic acid, β-hydroxybutyric acid, and acetone.

This denomination is not strictly accurate, because although acetoacetic acid and acetone are ketones, β-hydroxybutyric acid does not have a ketone group:

$CH_3\cdot CO\cdot CH_2\cdot COOH$	Acetoacetic acid
$CH_3\cdot CH\cdot OH\cdot CH_2\cdot COOH$	β-Hydroxybutyric acid
$CH_3\cdot CO\cdot CH_3$	Acetone

The term is nevertheless in common usage because variations of β-hydroxybutyric acid in the organism are parallel to those of acetone and acetoacetic acid.

Acetoacetic acid is the first substance formed. It is readily converted by reduction into β-hydroxybutyric acid. This process is activated by a specific enzyme, β-hydroxybutyric dehydrogenase, and a coenzyme, diphosphopyridinenucleotide (see Chap. 34).

Acetone is formed by decarboxylation of acetoacetic acid, a process that takes place without enzymes. Acetone excreted in the urine and the breath has a typical smell.

Embden and Marx[1] showed that a liver perfused with fatty acids containing an even number of carbon atoms produced a large amount of acetoacetic acid. The same result was obtained when slices of surviving liver were used.[2] This method is more precise than the perfusion method, and it has revealed that fatty acids with an odd number of carbon atoms also give rise to acetoacetic acid, but in much smaller quantities than fatty acids with an even number of carbon atoms.

Experiments with slices of surviving tissue also show that fatty acids with an even number of carbon atoms give rise to a larger amount of acetoacetic acid than might be expected if from

[1] KNOOP, F., *Beitr. z. chem. Physiol. u. Path.*, **6**, 150, 1904.

[1] EMBDEN, G., and A. MARX, *Beitr. z. chem. Physiol. u. Path.*, **11**, 318, 1908.

[2] QUASTEL, J. H., and A. H. M. WHEATLEY, *Biochem. J.*, **27**, 1753, 1933.

each molecule of fatty acid one molecule of acetoacetic acid were formed. It is therefore necessary to admit that the two-carbon-atom fragments broken off at each β-oxidation can combine with each other or with acetic acid and form acetoacetic acid. The β-oxidation theory postulates that the shortening of the fatty acid chain down to a fragment with four carbon atoms, which on oxidation will give acetoacetic acid, is the only source of this substance. This postulate can now no longer be accepted, and the β-oxidation theory must be slightly modified.

Recent work has shown that β-oxidation and condensation of the two-carbon-atom fragments take place.[1] Thus, palmitic acid (16 carbon atoms) splits into eight molecules of a two-carbon-atom substance, which on combining give four molecules of acetoacetic acid.

Enzyme systems. Semipurified enzymes that can oxidize fatty acids have been prepared. The system is a complex one; among the many factors which form it, adenosinephosphate, Mg^{++}, coenzyme I, and cytochrome c have been found. Oxidation is also accelerated by the addition of any one of the components of the tricarboxylic cycle, *e.g.*, fumaric, succinic, α-ketoglutaric acids, etc.[2] Oxidation of acetoacetic acid also takes place by the tricarboxylic cycle.

Experiments with isotopes. The utilization of isotopes to label substances has shown that fat stores are not static and inert accumulations, but rather that they are the site of a continuous interchange of substances. Fatty acids with a chain of less than 10 carbon atoms are not stored; they are either burned immediately or utilized in the synthesis of higher fatty acids. Other fatty acids are mostly stored and then gradually mobilized and burned. Fatty acids are changing continuously: they are dehydrogenated, disintegrated, lengthened, converted into other fatty acids, etc. Thus stearic acid can be converted into oleic or palmitic acid and vice versa.[3]

This method has also demonstrated how in the course of utilization of fatty acids their carbon atoms enter into the formation of ketone bodies and the components of the tricarboxylic cycle.

[1] MACKAY, E. M., *et al.*, *J. Biol. Chem.*, **135**, 157, 1940; **136**, 503, 1940; *J. Clin. Endocrinol.*, **3**, 101, 1943.

[2] MUÑOZ, J. M., and L. F. LELOIR, *J. Biol. Chem.*, **147**, 355, 1943; LEHNINGER, A. L., *J. Biol. Chem.*, **161**, 437, 1945.

[3] SCHOENHEIMER, R., and D. RITTENBERG, *J. Biol Chem.*, **120**, 155, 1937.

These few examples show how the use of isotopes has given a solid base to our knowledge of the mechanism of fatty-acid oxidation.

The mechanism of fatty-acid oxidation. The process of fatty-acid oxidation is not yet known with any certainty. A plausible, but by no means final, interpretation of the facts is that oxidation proceeds along the path described by Knoop, *i.e.*, β-oxidation with release of 2-carbon fragments. The substrate oxidized would not, however, be free fatty acid but a compound of the latter with coenzyme A (acetyl-CoA). The 2-carbon fragment broken off would not be acetic acid but acetyl-CoA. Acetyl would then be oxidized in the tricarboxylic cycle or would combine with another acetyl to form acetoacetic acid (see tricarboxylic cycle page 319).

KETOGENESIS AND KETOLYSIS

Excessive accumulation of ketone bodies is called ketosis. In this condition the concentration of ketone bodies in the blood rises above the normal level, and ketonuria increases. Ketosis leads to acidosis; therefore it causes a decrease in the alkali reserve in plasma, hyperventilation and a low CO_2 tension in alveolar air, increase in urinary ammonia, and in advanced stages a low blood pH (uncompensated acidosis). In severe acidosis, with an alkali reserve of 25, 20, or less, the subject becomes comatose or is on the verge of coma (*e.g.*, diabetic coma) and dies if adequate treatment is not given.

In normal conditions ketone bodies are formed in the liver and are oxidized in the tissues. When an excessive amount of fat is catabolized, the organism cannot burn it at a sufficiently rapid rate and ketone bodies accumulate, causing ketosis. Excessive catabolism of fats is observed when the diet does not contain a sufficient amount of carbohydrate, or when the organism cannot utilize carbohydrate at the normal rate, as in diabetes (Hirschfeld, 1895). In these cases the tissues utilize fats instead of carbohydrate.

An increase of ketone bodies is observed in the following circumstances: (*a*) in complete fasting; (*b*) when there is excess fat and insufficient carbohydrate in the diet; (*c*) in severe diabetes; (*d*) following injection of anterior hypophyseal extract.

The liver is the principal, perhaps the only, organ where ketone bodies are formed, as is

demonstrated by the following observations: (*a*) a liver perfused with fatty acids and some amino acids forms ketone bodies (Embden); (*b*) when fatty acids are added to slices of surviving liver, ketone bodies form, but in similar conditions slices of other tissues do not produce ketone bodies; (*c*) direct analysis of the liver reveals a higher concentration of ketone bodies than in any other tissue;[1] (*d*) these bodies are found in higher concentration in blood leaving the liver than in blood entering it;[2] (*e*) in hepatectomized animals pancreatectomy[3] and the injection of anterior hypophyseal extract[4] are not followed by an increase in ketone bodies.

Ketone bodies formed in the liver are oxidized in the tissues, especially in the muscles, in both normal and diabetic animals. This can be demonstrated in eviscerated animals[5] and *in vitro*, adding ketone bodies to minced muscle tissue.[6] The heart also utilizes ketone bodies. Administration of glucose and insulin diminishes the formation of ketone bodies (ketogenesis) but does not modify their oxidation by the tissues (ketolysis).

Administration of carbohydrate diminishes ketonuria provoked by fasting or an excessively fat diet. Insulin rapidly diminishes ketonuria in diabetes, because it acts on the liver, increasing the utilization of carbohydrate and diminishing that of fat and therefore the formation of ketone bodies.

Up to quite recently the aphorism "fats burn in the flame of carbohydrates" (Rosenfeld, 1906) was accepted as stating a well-proved fact. Later it was shown that the tissues of normal and diabetic animals utilize ketone bodies, a process that is not modified by the addition of glucose or insulin; therefore carbohydrate utilization does not increase ketolysis (destruction of acetoacetic acid) but diminishes hepatic ketogenesis. According to Stadie, acetoacetic acid is completely burned as long as the utilization of fats does not exceed a certain rate and no ketone bodies are excreted, but if fats are utilized at a higher rate, the amount of ketone bodies exceeds the capacity of the tissue to burn them, and the excess is accumulated and excreted into the urine. It is usually accepted now that the liver utilizes carbohydrate in preference to fat, so normally the amount of ketone bodies formed does not exceed the capacity of the tissues to burn them, and there is little or no ketonuria. When the liver has not enough carbohydrate at its disposal, or when its power of utilizing carbohydrate is diminished, it oxidizes fats and forms ketone bodies, thus increasing ketonemia and ketonuria.

Extirpation of the hypophysis or of the adrenals diminishes ketonuria in normal and diabetic animals, probably because these operations diminish protein and fat catabolism. Hyperthyroidism, on the contrary, increases ketogenesis in carbohydrate fasting and in diabetes. Anterior hypophyseal extract and some of the corticoadrenal hormones increase ketogenesis and ketonuria.

The Metabolism of Sterols

The metabolism of fats and phospholipids is closely related to that of sterols, which are not lipids. Cholesterol is an alcohol in which the hydroxyl is combined to a cyclopentanoperhydrophenathrene ring, which has side chains.

The term "steroid" is used for a large group of substances chemically related to the sterols by the nature of the constituent nucleus. This group includes, beside the sterols, the D vitamins, bile acids, sexual and corticoadrenal hormones, saponins, the glucosides of digitalis, substances in toad venom, carcinogenic substances, some of the "organizers" in the embryo, etc. These substances are of great biologic importance.

Sterols are essential constituents of cells. Apparently there is a chemical and functional equilibrium between sterols and lipids. There are also sterols in the blood, bile, and other organic fluids. They exist in the free state and combined with fatty acids as esters; usually there is a fixed ratio between these two states. The sterols are in the free state in bile, and in the erythrocytes and brain tissue they are mainly in the free state. In blood plasma, on the other hand, more than half the sterols are combined in esters

The sterol most commonly found in the animal organism is cholesterol, which on losing hydrogen acquires a second double bond and is converted into dehydrocholesterol (a provitamin

[1] Harrison, H. C., and C. N. H. Long, *J. Biol. Chem.*, **133**, 209, 1940.

[2] Crandall, J. A., *J. Biol. Chem.*, **135**, 139, 1940.

[3] Chaikoff, I. L., and S. Soskin, *Am. J. Physiol.*, **87**, 58, 1928.

[4] Mirsky, I. A., *et al.*, *Am. J. Physiol.*, **116**, 322, 1936.

[5] *Ibid.*, **118**, 290, 1937.

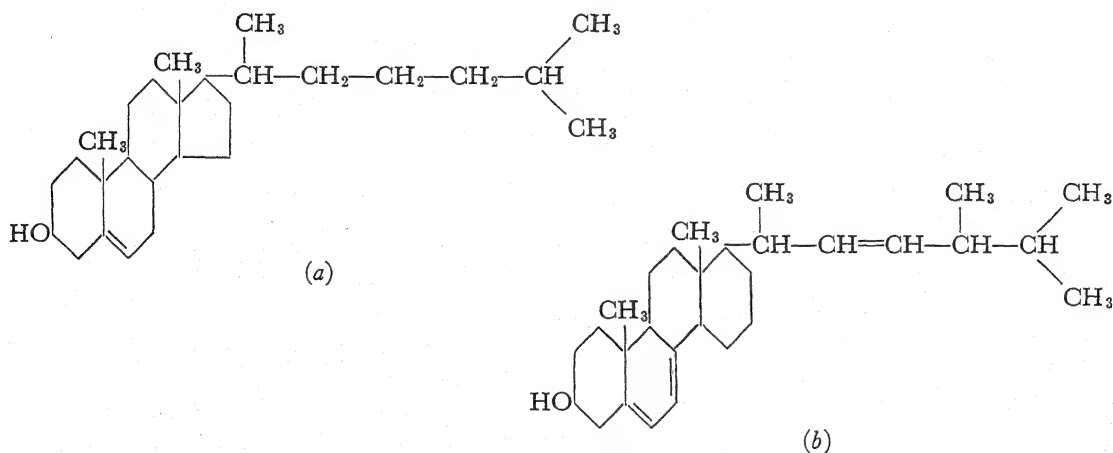
[6] Stadie, W. C., *et al.*, *J. Biol. Chem.*, **132**, 423, 1940; *J. Clin. Investigation*, **19**, 843, 1940; *Harvey Lect.*, **37**, 129, 1942.

which exists in the tissues) or into coprosterol found in the feces. Stigmasterol is the most abundant sterol in plants and ergosterol in fungi.

Absorption. Cholesterol and the D vitamins are the only sterols readily absorbed from the intestine. Cholesterol is not well absorbed unless fat is being absorbed simultaneously. Plant sterols, including ergosterol, are absorbed only in very small amounts. Bile, through its hydrotropic effect, helps the absorption of the D vitamins and probably also of cholesterol. Free cholesterol and cholesterol esters increase in the erythrocytes and blood plasma during absorption.

Nevertheless cholesterol synthesis proceeds at about one-third the rate of fatty-acid synthesis. A diet with abundant fat increases the synthesis of cholesterol.[1] Acetic acid is utilized in the synthesis of cholesterol by surviving slices of liver[2] when it is added to the substrate directly or can be split off from pyruvate, isovalerate, hexanate, octanate, or acetylacetate.

Intermediary metabolism and destruction. The intermediary metabolism of sterols is not well known. Generally, parallel changes in the concentration of sterols and fat in the blood and tissues are observed, but in some cases fat and sterols vary independently.

(*a*) (*b*)

Sterols: *a*, cholesterol; *b*, ergosterol.

Origin. Cholesterol has an endogenous and exogenous origin. Exogenous cholesterol is ingested with foods, particularly in eggs, brains, and liver. Endogenous cholesterol is the product of synthesis by the tissues. The existence of this synthesis has been demonstrated repeatedly, but its chemical process is still unknown. Larger amounts of cholesterol are found by measuring the cholesterol content of the body than can be accounted for by the cholesterol ingested[1] in animals fed on a diet poor in cholesterol. In mice fed a cholesterol-free diet and "heavy" water (*i.e.*, water in which hydrogen is replaced by deuterium), cholesterol with deuterium in its molecule has been found; an amount of cholesterol equivalent to that existing in the whole body of the animal at the beginning of the experiment can be synthesized in 1 month.[2]

[1] Channon, H. S., *Biochem. J.*, **19**, 424, 1925.

[2] Rittenberg, D., and R. Schoenheimer, *J. Biol. Chem.*, **121**, 235, 1937.

Destruction of cholesterol in the organism has been demonstrated in birds, rabbits, and mice. A mouse fed on a diet containing large quantities of cholesterol can destroy in 1 month five times the amount existing in the whole body at the beginning of the experiment.[3] The destruction of sexual steroids by the liver has been well demonstrated and some of the end products are known; *e.g.*, some of the testicular and corticoadrenal steroids give rise to 17-ketosteroids which are excreted in the urine.

Excretion. Cholesterol and its derivatives are excreted (*a*) in the bile, as free cholesterol, part of which is reabsorbed from the intestine; (*b*) through the mucosa of the large intestine,

[1] Eckstein, H. C., *J. Biol. Chem.*, **125**, 99 and 107, 1938.

[2] Bloch *et al.*, *J. Biol. Chem.*, **162**, 441, 1946; **107**, 811, 1948.

[3] Schoenheimer, R., and F. Breusch, *J. Biol. Chem.*, **103**, 439, 1933.

which excretes dehydrocholesterol and coprosterol. There is no cholesterol in normal urine, but it is found in cases in which there is infiltration of the renal tubes by cholesterol deposits (lipoid nephrosis).

Functions. Cholesterol seems to play some part in the following processes: (*a*) in the absorption and transport of fats; (*b*) in the permeability of cell membranes (therefore, in the interchange of water and substances in solution); (*c*) in the isolation of the nerve impulse (cholesterol in the neural sheaths); (*d*) in the neutralization of hemolytic and cytolytic substances, such as saponins and phospholipid derivatives (lysocitin and lysocephalin)—an effect that has fostered the mistaken belief that it has universal antitoxic properties. Cholesterol is a precursor in the synthesis of bile acids and of sexual and corticoadrenal hormones. Dehydrocholesterol and irradiated ergosterol are converted into the antirachitic D vitamins.

A diet containing an excess of cholesterol provokes in certain species a fatty liver in which the formation and destruction of phospholipids are diminished. An excess of biotin in the diet produces infiltration of fat and cholesterol in the liver, which can be prevented by an adequate dose of inositol.

In the rabbit (Anitschkow) and guinea pig and in birds, cholesterol feeding provokes hypercholesterolemia with deposition of cholesterol in the arterial walls, which show lesions similar to those of atherosclerosis. Giant lipoprotein molecules appear in the serum; they have a high molecular weight, but owing to their large fat component they have a low specific gravity. They can be separated by fractionated ultracentrifugation, and the "atherosclerogenous band" (Sf 12 to 100) differs from the normal (Sf 10 or less) (Gofman *et al.*).

Lipoid thesaurismosis (lipoidosis). Abnormal storage of fat is observed in several diseases. Cholesterol and its esters are deposited in the following conditions: (*a*) localized or generalized xanthomatosis: (*b*) diabetic xanthomatosis; (*c*) Hand-Schüller-Christian disease, in which yellow nodules are formed in the bones. In this last condition there is also polyuria, exophthalmos, etc. Deposition of phospholipids in the spleen, liver, and brain is observed in Niemann-Pick disease; sphyngomyelin is particularly increased. Cerebrosides are accumulated in Gaucher's disease, in which the spleen is enlarged, there is thrombopenia, and radiographs show maculae in the bones; large amounts of kerasin are formed in this disease.

BIBLIOGRAPHY

Best, C. H., and C. C. Lucas, Choline. Chemistry and Significance as a Dietary Factor, *Vitamins & Hormones*, **1**, 1, 1943.

Bills, C. E., Physiology of the Sterols, *Physiol. Rev.*, **15**, 1, 1935.

Bloor, W. R., "Biochemistry of the Fatty Acids," Reinhold, New York, 1943.

Brown, J. B., The Chemistry of Lipids, *Ann. Rev. Biochem.*, **15**, 93, 1946.

Burr, G. O., and R. H. Barnes, Non-caloric Functions of Dietary Fats, *Physiol. Rev.*, **23**, 257, 1943.

Chaikoff, J. L., The Application of Labelling Agents to the Study of Phospholipid Metabolism, *Physiol. Rev.*, **22**, 291, 1942.

Deuel, H. J., "Chemistry and Biochemistry of the Lipids," Interscience Publishers, Inc., New York, 1951.

Frazer, A. C., Fat Absorption and Its Relation to Fat Metabolism, *Physiol. Rev.*, **20**, 560, 1940.

Gofman, J., *et al.*, Lipoproteins and Atherosclerosis, *Circulation*, **5**, 119, 1952.

Gurin, S., and R. O. Brady, Cholesterol Synthesis, "Ciba Foundation Conference on Isotopes in Biochemistry," J. & A. Churchill, London, 1951.

Jukes, T., Choline, *Ann. Rev. Biochem.*, **16**, 193, 1947.

Leathes, J. B., and H. S. Raper, "The Fats," Longmans, New York, 1925.

Lehninger, A. L., On the Activation of Fatty Acid Oxidation, *J. Biol. Chem.*, **161**, 437, 1945.

Leloir, L. F., The Mechanism of Fatty Acid Oxidation, *Enzymologia*, **12**, 263, 1948.

McHenry, E. W., and J. M. Patterson, Lipotropic Factors, *Physiol. Rev.*, **24**, 128, 1944.

Piskur, M. M., and H. W. Schultz, The Chemistry and Metabolism of Lipids, *Ann. Rev. Biochem.*, **16**, 79, 1947.

Reichstein, T., and H. Reich, The Chemistry of Steroids, *Ann. Rev. Biochem.*, **15**, 155, 1946.

Rynearson, E. H., and C. F. Gastineau, "Obesity," Blackwell, Oxford, 1949.

Schoenheimer, R., and D. Rittenberg, The Study of the Intermediary Metabolism of Animals with the Aid of Isotopes, *Physiol. Rev.*, **20**, 218, 1940.

Sinclair, R. G., Fat Metabolism, *Ann. Rev. Biochem.*, **6**, 245, 1937.

Sinclair, R. G., The Anabolism and Function of the Phospholipids, *Biol. Symposia*, **5**, 82, 1941.

Stadie, W. C., Fat Metabolism, *Ann. Rev. Biochem.*, **15**, 219, 1946.

Stadie, W. C., The Intermediary Metabolism of Fatty Acids, *Physiol. Rev.*, **25**, 395, 1945.

Verzár, P., and E. J. McDougall, "Absorption from the Intestine," Longmans, New York, 1936.

CHAPTER 43

Protein Metabolism

CONSTITUTION OF PROTEINS

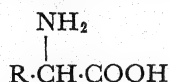
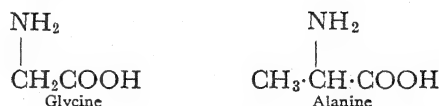
Protein is an essential constituent of protoplasm in all living organisms. It differs from carbohydrate and fat because it is made up not only of carbon, oxygen, and hydrogen, but also of nitrogen, sulfur, and in certain cases phosphorus. The average elementary composition of protein (globulins and albumins) is as follows: C, 54 per cent; H, 7 per cent; O, 22 per cent; N, 16 per cent; S, 1 per cent. Therefore 1 gm. of protein has 0.16 gm. of N, and 1 gm. of protein N is the equivalent of 6.25 gm. of protein. Conjugated proteins are those in whose molecule a nonprotein group, called in some cases the prosthetic group, is joined to a protein group; *e.g.*, hemoglobin, glycoproteins, lipoproteins, and nucleoproteins.

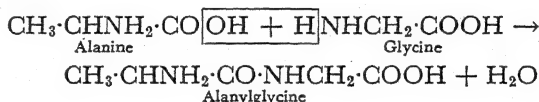
Proteins are made up of numerous units or "building stones," known as amino acids. Protein metabolism is mainly the metabolism of amino acids. Proteins are of great but variable molecular weight, *e.g.*, 66,400 for hemoglobin, 5,000,000 for hemocyanin.

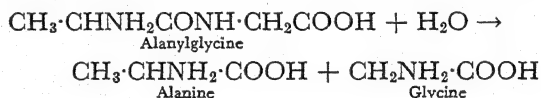
Natural amino acids are fatty acids in which one of the alpha hydrogen atoms has been replaced by an NH_2 group. Therefore an amino acid has a basic (NH_2) group and an acid (COOH) group. The simplest amino acid is glycine (also called glycocoll), which is aminoacetic acid. The next one is alanine (α-aminopropionic acid). With the exception of proline and oxyproline, in which there is an NH group in the ring, amino acids can be represented by formulas of the following type, in which R represents the rest of the amino-acid molecule:

$$\begin{array}{c} NH_2 \\ | \\ R \cdot CH \cdot COOH \end{array}$$

For example:

$$\underset{\text{Glycine}}{\begin{array}{l} NH_2 \\ | \\ CH_2COOH \end{array}} \qquad \underset{\text{Alanine}}{\begin{array}{c} NH_2 \\ | \\ CH_3 \cdot CH \cdot COOH \end{array}}$$

The basic group of one amino acid may combine with the acid group of another, with loss of water:

$$\underset{\text{Alanine}}{CH_3 \cdot CHNH_2 \cdot CO}\boxed{OH + H}\underset{\text{Glycine}}{NHCH_2 \cdot COOH} \rightarrow \underset{\text{Alanylglycine}}{CH_3 \cdot CHNH_2 \cdot CO \cdot NHCH_2 \cdot COOH} + H_2O$$

The union CO—NH is known as a peptide linkage. The reverse process, which also takes place in the peptide linkage, causes the separation of amino acids; it is a process of hydrolysis.

$$\underset{\text{Alanylglycine}}{CH_3 \cdot CHNH_2CONH \cdot CH_2COOH} + H_2O \rightarrow \underset{\text{Alanine}}{CH_3 \cdot CHNH_2 \cdot COOH} + \underset{\text{Glycine}}{CH_2NH_2 \cdot COOH}$$

In theory the natural amino acids could form an enormous number of proteins, but there is evidence that this is not so. Amino acids seem to be distributed in the polypeptide chain in a certain order peculiar to each protein.

Theoretically there are two stereoisomers for each amino acid, but with the exception of threonine, all natural amino acids are the L and not the D isomers.

The architecture of the protein molecule has been studied by means of x-ray diffraction photographs (Astbury), and several optical methods. Fibrous proteins, such as keratin, collagen, and myosin, have long threadlike molecules. These filaments can be stretched, and afterward recover their initial length. The mole-

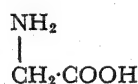
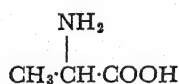
cule changes in length by processes of folding and unfolding, similar to a concertina when it is being played. Globular proteins, *e.g.*, blood plasma proteins and egg albumen, are formed by chains or filaments wound in a compact mass.

Proteins form colloidal solutions because of the large size of their molecules. Proteins and amino acids are ampholytes, because the acid group (COOH) can set free H^+ and the basic group (NH_2) can accept H^+ or set free $OH-$ from H_2O. In an acid medium proteins behave as bases, and in alkaline medium they behave as acids, forming salts (proteinates) such as sodium or potassium proteinate. At a certain hydrogen ion concentration of the solvent, the protein molecule is electrically neutral, its dissociation is at a minimum, and the concentration of acid ions is equal to that of basic ions. This is known as the isoelectric point. The charge, solubility, stability, and viscosity of the majority of proteins are at a minimum when they are in a solvent at the isoelectric point; some of them precipitate. Isoelectric precipitation is used to separate and purify proteins.

CLASSIFICATION OF THE AMINO ACIDS

I. Aliphatic amino acids

A. *Monoaminomonocarboxylic acids*

1. Glycine (or glycocoll) $C_2H_5NO_2$, aminoacetic acid:

$$\begin{array}{l} NH_2 \\ | \\ CH_2\cdot COOH \end{array}$$

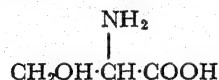
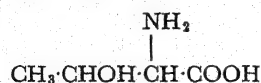
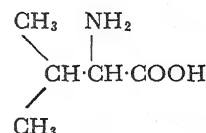
2. Alanine $C_3H_7NO_2$, α-aminopropionic acid:

$$\begin{array}{l} \quad\ NH_2 \\ \quad\ \ | \\ CH_3\cdot CH\cdot COOH \end{array}$$

3. Serine $C_3H_7NO_3$, β-hydroxy-α-aminopropionic acid.

$$\begin{array}{l} \qquad\quad NH_2 \\ \qquad\quad\ | \\ CH_2OH\cdot CH\cdot COOH \end{array}$$

4. Threonine $C_4H_9NO_3$, α-amino-β-hydroxybutyric acid:

$$\begin{array}{l} \qquad\qquad\ NH_2 \\ \qquad\qquad\ \ | \\ CH_3\cdot CHOH\cdot CH\cdot COOH \end{array}$$

5. Valine $C_5H_{11}NO_2$, α-aminoisovalerianic acid:

$$\begin{array}{l} CH_3 \quad NH_2 \\ \quad \diagdown \quad\ | \\ \qquad CH\cdot CH\cdot COOH \\ \quad \diagup \\ CH_3 \end{array}$$

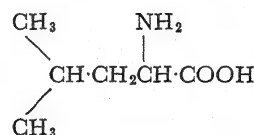
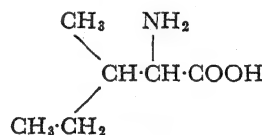
6. Norleucine $C_6H_{13}NO_2$, α-aminocaproic acid:

$$\begin{array}{r} NH_2 \qquad\quad \\ | \qquad\qquad \\ CH_3\cdot CH_2\cdot CH_2\cdot CH_2\cdot CH\cdot COOH \end{array}$$

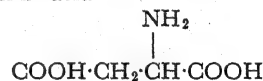
7. Leucine $C_6H_{13}NO_2$, α-aminoisocaproic acid:

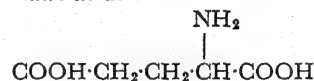
$$\begin{array}{l} CH_3 \qquad\quad NH_2 \\ \quad \diagdown \qquad\quad | \\ \qquad CH\cdot CH_2CH\cdot COOH \\ \quad \diagup \\ CH_3 \end{array}$$

8. Isoleucine $C_6H_{13}NO_2$, β-methyl-β-ethyl-α-aminopropionic acid:

$$\begin{array}{l} \quad CH_3 \quad NH_2 \\ \qquad \diagdown \quad\ | \\ \qquad\quad CH\cdot CH\cdot COOH \\ \qquad \diagup \\ CH_3\cdot CH_2 \end{array}$$

B. *Monoaminodicarboxylic acids*

9. Aspartic acid $C_4H_7NO_4$, α-aminosuccinic acid:

$$\begin{array}{r} NH_2 \qquad\quad \\ | \qquad\qquad \\ COOH\cdot CH_2\cdot CH\cdot COOH \end{array}$$

10. Glutamic acid $C_5H_9NO_4$, α-aminoglutaric acid:

$$\begin{array}{r} NH_2 \qquad\quad \\ | \qquad\qquad \\ COOH\cdot CH_2\cdot CH_2\cdot CH\cdot COOH \end{array}$$

C. *Diaminomonocarboxylic acids*

11. Arginine $C_6H_{14}N_4O_2$, δ-guanidine-α aminovalerianic acid:

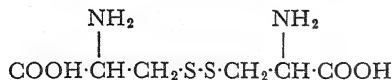
$$\begin{array}{l} \qquad NH_2 \qquad\qquad\qquad\qquad NH_2 \\ \qquad\ | \qquad\qquad\qquad\qquad\qquad | \\ NH{=}C\cdot NH\cdot CH_2\cdot CH_2\cdot CH_2\cdot CH\cdot COOH \end{array}$$

12. Lysine $C_6H_{14}N_2O_2$, α-ϵ-diaminocaproic acid:

$$\begin{array}{l} NH_2 \qquad\qquad\qquad\quad NH_2 \\ | \qquad\qquad\qquad\qquad\quad | \\ CH\cdot CH_2\cdot CH_2\cdot CH_2\cdot CH\cdot COOH \end{array}$$

D. *Sulfur-containing amino acids*

13. Cystine (or dicysteine) $C_6H_{12}N_2S_2O_4$, di-β-thio-α-aminopropionic acid:

```
    NH2                NH2
    |                  |
COOH·CH·CH2·S·S·CH2·CH·COOH
```

14. Methionine $C_5H_{11}SNO_2$, α-amino-γ-methylthio-*n*-butyric acid:

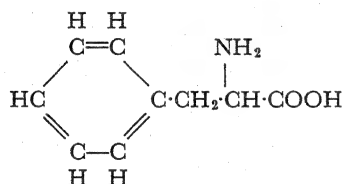
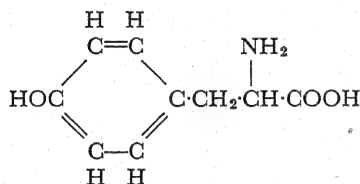
```
                  NH2
                  |
CH3—S—CH2·CH2·CH·COOH
```

II. Aromatic amino acids

15. Phenylalanine $C_9H_{11}NO_2$, β-phenyl-α-aminopropionic acid:

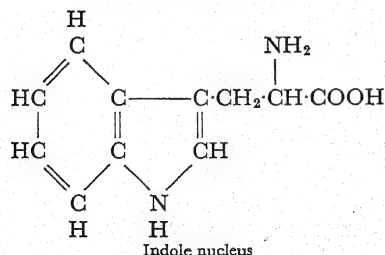
```
    H  H
    C=C         NH2
   /   \        |
HC      C·CH2·CH·COOH
   \\  //
    C—C
    H  H
```

16. Tyrosine $C_9H_{11}NO_3$, β-parahydroxyphenyl-α-propionic acid:

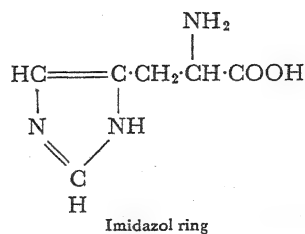
```
     H  H
     C=C         NH2
    /   \        |
HOC      C·CH2·CH·COOH
    \\  //
     C—C
     H  H
```

III. Heterocyclic amino acids

17. Tryptophane $C_{11}H_{12}N_2O_2$, β-indole-α-aminopropionic acid:

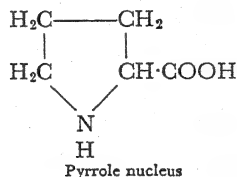
```
    H
    C              NH2
  //  \            |
HC     C——————C·CH2·CH·COOH
 |     ||     ||
HC     C      CH
  \\  /  \   /
    C      N
    H      H
```

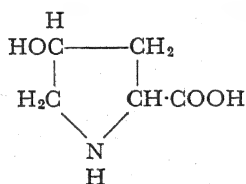
Indole nucleus

18. Histidine $C_6H_9N_3O_2$, α-amino-β-imidazolpropionic acid:

```
                 NH2
                 |
HC════C·CH2·CH·COOH
|       |
N       NH
  \\   /
    C
    H
```

Imidazol ring

19. Proline $C_5H_9NO_2$, pyrrolidinecarboxylic acid:

```
H2C——————CH2
 |        |
H2C      CH·COOH
   \    /
     N
     H
```

Pyrrole nucleus

20. Hydroxyproline (or oxyproline) $C_5H_9NO_3$, γ-hydroxy-α-pyrrolidinecarboxylic acid:

```
  H
HOC——————CH2
 |        |
H2C      CH·COOH
   \    /
     N
     H
```

This list of amino acids found in animal organisms is completed by the addition of 3-5-diiodotyrosine (or iodogorgoric acid), thyroxine, citruline, ornithine, and a few others.

ABSORPTION OF PROTEIN

The process of digestion hydrolyzes proteins into the constituent amino acids, which are then absorbed in the small intestine. Amino acids are simple diffusible substances, the "building stones" that the organism will use to reconstruct its own large specific protein molecules. Proteins have the ability to produce immunologic reactions that are species-specific. Amino acids do not produce any immunologic reactions.

Native proteins are not usually absorbed. In certain conditions, however, particularly in young animals, very small amounts can be absorbed. This should be considered more a deficiency in digestion than a normal process. When a protein from another species enters into the circulation, it acts as a foreign body and pro-

vokes the formation of antibodies. The organism becomes sensitized; after one to three weeks it is allergic to the protein, and may respond with an anaphylactic shock to another dose of it.

Specific precipitins for the protein in question are usually found in the subject's serum. Foreign proteins are toxic; the reactions they provoke are usually transitory, but if the administration is prolonged cachexia and death may result (protein intoxication). In man, injection of blood serum from another species frequently causes what is known as serum sickness; this begins a few days after the injection with a nettle rash (urticaria) or cutaneous erythema, painful joints, edema, fever, etc. An acquired capacity to react to foreign proteins is revealed by food idiosyncrasies, hay fever, asthma, and other abnormal reactions to protein in food, fungi, pollen, etc.

Several workers have reported the absorption of small amounts of polypeptides in the course of digestion,[1] but only insignificant amounts of protein food are absorbed in this form. Nearly all protein is absorbed in the form of amino acids. Blood normally contains amino acids (5 to 8 mg. of amino-acid nitrogen in 100 ml. of blood, in fasting subjects), which are in a free state, as is demonstrated by dialyzing circulating blood (Abel's vividiffusion technique) and by paper chromatography (Dent). In the tissues they are found in a slightly higher concentration. In the course of digestion, amino-acid concentration increases in the blood of the portal and jugular veins,[2] especially essential amino acids of high molecular weight. At the same time there is an increase in the urinary excretion of amino acids, especially of nonessential amino acids of low molecular weight. Human blood plasma given by mouth is absorbed and its nitrogen is utilized by the organism; its absorption, however, is not accompanied by an increase in amino acids in the blood.

Amino acids in the tissues also increase during absorption, especially in the liver, muscles, and kidneys. The liver is the tissue that takes up amino acids at the highest rate and then rapidly deaminizes them, producing urea, which increases in the blood during absorption of protein.

[1] London, E. S., and N. Kotschneff, *Ztschr. f. physiol. Chem.*, **228**, 235, 1934; Godfried, *Biochem. J.*, **33**, 955, 1939.

[2] Van Slyke, D. D., and G. M. Mayer, *J. Biol. Chem.*, **16**, 197 and 213, 1913; Van Slyke, D. D., *Science*, **95**, 259. 1942; Dent, C. E., *Biochem. Symposia*, **3**, 25, 1949.

An important, although indirect, demonstration that amino acids are the nutritional basis of protein metabolism was given by Rose, when he showed that normal growth can be maintained by the administration of a mixture of 10 amino acids as the only source of nitrogen. That nitrogen equilibrium can be maintained by the ingestion[1] or injection[2] of abiuretic products of protein digestion is a fact that has been known for a long time.

PROTEIN BALANCE

Protein in food is broken down by digestion to amino acids, which are absorbed and utilized in the formation of the proteins specific for each animal and for each tissue (protein anabolism). At the same time protein and amino acids are being disintegrated and converted into other substances (protein catabolism). The greater part of the end products of protein catabolism is eliminated in the urine as nitrogenous substances; a small part is excreted in the feces. Fecal nitrogen is mostly nitrogen of foodstuffs that have not been absorbed. The rest, about 1 gm. per day (0.5 to 1.5 gm.), is nitrogen in the intestinal flora. Only a small part comes from the residue of digestive juices and cells shed by the intestinal mucosa. An insignificant amount of nitrogen is eliminated in the sweat—about 0.1 mg. daily (Benedict). Small quantities are lost when hair and nails are cut, in cutaneous desquamation, and in other secretions. Gaseous nitrogen is not fixed by the animal organism, and it plays no part in its metabolism.

If the intake of nitrogen is greater than the output, nitrogen is gained and it is said there is a positive balance. This occurs during growth, in pregnancy, in the development of muscle tissue by muscular exercise, and in convalescence after wasting disease. If the output of nitrogen is greater than the intake, there is loss of nitrogen and it is said there is a negative balance. This occurs in undernourishment, inanition, and cases of excess protein catabolism such as in fever, after accidental trauma or surgical operations, and in wasting diseases. In these cases nitrogen loss is compensated with difficulty; if

[1] Loewi, 1902; Henderson and Dean, 1903; Abderhalden *et al.*, 1907 to 1913. Enzymatic hydrolysis does not destroy the essential amino acids; acid hydrolysis destroys some of them.

[2] Henriques and Hansen, 1905; Henriques and Andersen. 1913.

the patient cannot ingest, digest, or absorb proteins, it is advisable to inject amino-acid solutions.[1] During the period of convalescence a positive nitrogen balance is established, and later the subject returns to nitrogen equilibrium.

A nursing mother may have a negative nitrogen balance because of nitrogenous substances secreted in the milk, but she maintains her own body proteins intact if she is adequately fed.

The adult organism in nitrogen equilibrium utilizes protein in preference to other substances, but it is not stored as is the case with fat and carbohydrate. In the 12 to 24 hr. following the administration of protein or amino acids, an equivalent amount of nitrogen is excreted in the urine. If the subject receives 10 gm. of protein nitrogen, he will eliminate 10 gm. of nitrogen in the urine; if 20 gm. is given him, 20 gm. will be eliminated. Nitrogen is retained in the organism only when there is a positive nitrogen balance, *i.e.*, in the circumstances mentioned in the preceding paragraph.

Observation of protein balance shows the nutritive value of the different proteins in the maintenance of nitrogen equilibrium, growth, and the repair of tissues.

PROTEIN REQUIREMENT

The protein in food is needed for growth and repair of tissues and for the maintenance of body proteins and body weight, because protein is being continuously catabolized. In the course of protein catabolism, for every gram of protein converted in the organism, 4.1 kg.-cal. is set free.

Protein deficiency delays or stops growth and causes loss of weight. The normal wear and tear of tissues is not repaired, *e.g.*, blood plasma and hemoglobin are not replaced. The healing of wounds is retarded, and resistance to trauma and infection is diminished. Fatty degeneration and cirrhosis of the liver can be prevented and sometimes cured by the administration of a sufficient amount of protein.

When all food is withheld (total fasting) or only protein is eliminated from an otherwise complete diet (protein fasting), the destruction of body protein does not cease, and nitrogenous end products of protein catabolism are still eliminated in the urine and feces. Nitrogen excretion diminishes, nevertheless, and after a 31- to 38-day fast it has been observed to fall to 6.9 and 3.3 gm. per day, whereas a normal, well-fed adult eliminates from 12 to 16 gm. daily. In protein fasting less nitrogen is eliminated than in total fasting; ingestion of carbohydrate and fat diminishes the amount of protein metabolized, and excretion of nitrogen is reduced to 3.8 gm. and even to 1.75 gm. daily. Protein is thus spared by the utilization of other nutritive substances. Protein sparing is greater if carbohydrate alone is ingested than if fat alone is ingested; but maximum protein sparing is obtained by the ingestion of carbohydrate and fat in amounts of equal caloric value.

Some hormones (hypophyseal, thyroid, testicular, etc.) are necessary during certain periods of life to maintain growth and protein anabolism. Extirpation of the anterior hypophysis or of the thyroid stops growth, and the formation of protein ceases almost completely. The administration of anterior hypophyseal extract or the growth hormone (somatotrophin), and of testosterone provokes the retention of nitrogen and increases the formation of protein. On the other hand excess thyroid secretion and some of the corticoadrenal hormones increase protein catabolism and therefore nitrogen excretion (see Chap. 53, The Thyroid Gland, and Chap. 54, The Adrenal Glands).

The minimum physiologic protein requirement is the smallest amount of protein needed to maintain nitrogen equilibrium. To determine this minimum the subject is given a protein-free but otherwise complete diet, and then increasing quantities of protein are given until nitrogen equilibrium is obtained. The minimum varies in different subjects and according to the experimental conditions. Figures given in the numerous publications on the subject vary from ⅓ to ⅔ gm. of protein per kilogram of body weight per day. The average of several experiments by different workers has been given at 0.0548 gm. of N, corresponding to 0.34 gm. of protein per kilogram per day.[1] Sherman gives figures from 21 to 65 gm. (average 44.4 gm.) of protein per day, which corresponds approximately to 0.66 gm. per kg. per day.

According to Berg[2] many factors influence protein catabolism. In the first place habit is important; thus

[1] Peters, J. P., *et al.*, Seminar on Protein Hydrolysates, *Am. J. Med.*, **5**, 1948.

[1] Bertram, F., and A. Bornstein, in Bethe's "Handbuch der normalen und pathologischen Physiologie," **5**, 84, Springer, Berlin, 1928.

[2] Berg, O., "Eiweissbedarf und Mineralstoffwechsel bei einfachster Ernährung," S. Hirzel, Leipzig, 1931.

a herbivorous subject utilizes a vegetable diet, and a carnivorous one a meat diet, better than other diets, owing to a special adaptation of the digestive tract and metabolism. Nitrogen excretion is dependent on the ratio of base to mineral acids formed by the metabolization of the diet. If there is an excess of acid end products and the urine is acid, more nitrogen is excreted than if base is predominant. Base is ingested in the food, and mineral acids are mostly end products of protein metabolism.

The standard minimum protein requirement is the smallest amount of protein needed to maintain optimum conditions of health. It is very difficult to say what this minimum is for man. Careful observations made on human subjects have shown that it is possible to keep in excellent health with small amounts, *e.g.*, 40 to 50 gm. daily (Chittenden), 35 gm. (Sherman), and even 22 gm. (Hinhede). It is usually accepted that protein intake should not fall below 1 gm. per kg. of body weight per day; half the protein should be of first-class biologic value, *i.e.*, animal protein (milk, eggs, meat, glands such as liver and kidney). These amounts are probably more than twice the minimum physiologic protein requirement of the majority of individuals. They assure, however, a margin of safety, especially important in the prevention of insufficiencies due to the ingestion of poor-quality protein.

Man can live in excellent health on an almost exclusively meat diet. Krogh has observed a 35- to 85-gm. daily nitrogen excretion in Eskimos (this corresponds to a daily catabolism of up to 530 gm. protein). Argentine cattle drovers were accustomed to eat regularly 2 to 3 kg. of meat daily.

In recent years, the great importance of protein deficiency in the diet has been well demonstrated and understood. The effects of protein deficiency and the means of preventing it have been carefully studied. It has also been shown that men fed a high-protein diet are of greater height and weight than others, of the same race, fed on a lower protein diet.

Protein deficiency is due to (*a*) insufficient protein in the diet, with respect to either quantity or quality; (*b*) disturbances in digestion or absorption of protein (*e.g.*, enteritis, diarrhea, etc.); (*c*) insufficient protein synthesis by the organism (*e.g.*, certain forms of hepatic insufficiency); (*d*) increase in protein requirement not covered by the diet (growth, pregnancy, lactation); (*e*) excess protein destruction due to trauma or disease (burns, accidental trauma, surgical operations, fever, severe diabetes, wasting disease, infections, etc.); (*f*) abnormal loss of protein (burns, albuminuria, hemorrhage, frequent evacuation of peritoneal or pleural effusions).

Loss of protein is made evident by a negative nitrogen balance. There is loss of body weight. Resistance to infection and trauma diminishes, and healing is delayed. The liver is easily damaged, and hepatic insufficiency occurs. There is hypoproteinemia, due especially to a decrease in serum albumin. Marked hypoproteinemia is accompanied by edema. Elman has calculated that for each gram of plasma protein lost there is a corresponding loss of about 30 gm. of tissue protein.

Protein deficiency is remedied by increasing the protein in the diet or, when protein cannot be digested, by injecting protein hydrolysate. In certain serious preoperative or postoperative conditions and in cases of severe malnutrition, protein equilibrium has been maintained by transfusion of blood or plasma, this being the only source of nitrogen. Nitrogen equilibrium has also been maintained by injecting solutions of amino acids containing the "essential" amino acids or of casein hydrolysate.[1]

The maintenance of nitrogen equilibrium for a short time does not signify that protein is being synthesized. For this to occur, nitrogen, phosphorus, and potassium must be accumulated, and the increase in weight observed must not be due solely to retention of water.[2]

THE BIOLOGIC VALUE OF PROTEINS

The biologic value of a protein depends on (*a*) its digestibility; (*b*) its adequacy in providing material for the formation of body protein necessary for the maintenance of nitrogen equilibrium or for growth.

The digestibility of a protein is measured by the proportion of the amount ingested that is absorbed. This measurement is carried out as follows: The amount of nitrogen in the food is determined. Then the fecal nitrogen is deter-

[1] ELMAN, R., *J. A. M. A.*, **128**, 659, 1945; *Advances in Protein Chemistry*, **3**, 269, 1947; "Parenteral Alimentation in Surgery with Protein and Amino Acids," Hoeber, New York, 1947.

[2] MOORE, F. D., *Nutrition Rev.*, **6**, 161. 1948.

mined and subtracted from the former.[1] The difference corresponds to nitrogen absorbed. Thus if, out of 100 gm. of ingested protein, 95 gm. is absorbed, digestibility is said to be 95 per cent. Digestibility of animal proteins is above 95 per cent; that of potatoes and other vegetables is about 80 per cent.

Table 49. Biologic Value of Proteins in Man Compared with Their Chemical Value

Source of protein	Chemical value*	*Biologic value*			
		Thomas, 1929	*Murlin, Edwards, and Hawley, 1944*	*Mitchell and Savhun, 1948*	*Other workers*
Whole egg	100	..	97	78	65
Fish	72	95			
Beef	71	105	84	72	85
Milk	68	100	..	74	62
Rice	44	88			
Potato	..	79			
Whole wheat	37	..	..	55	33
Spinach	..	64			
Corn flour	28	30	..	45	
White flour	28	40	..	41	

Source: Block, R. J., and H. H. Mitchell, *Nutrition Abstr. & Rev.*, **16**, 249, 1946; Mitchell, H. H., and Sayhun, M., "Proteins and Amino Acids in Nutrition," Reinhold, New York, 1948.

* Percentage of essential amino acids.

The biologic value of a protein is determined by giving the subiect a protein-free, but otherwise complete, diet, to which the protein to be studied is added in variable amounts (Osborne and Mendel). The growth value is usually expressed in grams of weight increase per gram of protein given. The maintenance value is expressed by the minimum amount ingested (per gram of body weight per week) that maintains a stable body weight. Mitchell expresses the maintenance or growth value of a protein by the percentage of nitrogen retained in the body.

In man several means have been used to calculate the biologic value of proteins (Table 49).

[1] The fasting subject eliminates 0.5 to 1.5 gm. of nitrogen in the feces; therefore from the total fecal nitrogen 1 gm. is deducted.

Thomas[1] determined the amount of body protein spared, in human subjects on a protein-free diet, by the ingestion of the protein studied. Milk was considered the standard and given a value of 100; others take whole egg as the standard. He found that animal proteins are far superior to vegetable proteins in this respect. Other workers have not observed such great differences.

Complete and incomplete proteins. Osborne and Mendel's work[2] has shown there are some proteins by which growth can be maintained, others by which body weight is maintained but with which growth ceases, and others with which weight is lost and the animal cannot live. The two latter are known as incomplete proteins. Their deficiency in respect of nutritive properties is due to the fact that certain amino acids do not form part of their molecule. Thus gelatin does not contain tryptophan, cystine, valine, or phenylalanine; if these amino acids are added to a diet of which the only protein is gelatin, the animals grow normally. Another typical example is given by zein (the protein in corn), which does not contain tryptophane or lysine. Growth and weight are not maintained in animals when zein is the only protein in the diet. On adding tryptophane to the zein, the animals do not lose weight, but they do not grow. If both tryptophane and lysine are added to zein, a normal growth curve is obtained (Fig. 204).

Essential amino acids. Rose[3] and his associates have done important work on the amino acids needed for normal nutrition. They fed animals with various mixtures of the 22 amino acids that form part of natural proteins, these being the only source of nitrogen the animals had. They proved that ten amino acids were essential for normal growth in the rat. Twelve were not essential (Table 50). The essential ones are not synthesized in the organism at a sufficiently rapid rate for normal growth. Only eight of these amino acids seem to be indispensable in man.[4] Histidine and arginine are necessary for the rat, but it has not been definitely established that they are essential in human

[1] Thomas, K., *J. Nutrition*, **30**, 97, 1929.

[2] Osborne, T. B., and L. B. Mendel, *J. Biol. Chem.*, **17**, 325, 1914.

[3] Rose, W. C., *Physiol. Rev.*, **18**, 109, 1938.

[4] Rose, W. C., W. J. Haines, and J. E. Johnson, *J. Biol. Chem.*, **182**, 541, 1950; **188**, 49, 1951; **193**, 605, 1951; Rose, W. C., *Proc. Am. Philos. Soc.*, **91**, 113, 1947; *Federation Proc.*, **8**, 546, 1949.

nutrition.[1] Ten amino acids are not indispensable for the rat or man.

When an essential amino acid is omitted from the diet, there is a negative nitrogen balance, with loss of appetite, and the subject feels tired and irritable. On adding the amino acid to the diet, these signs and symptoms disappear. Observations in man have been made over only short periods, and the effect of amino-acid insufficiency on several functions, such as growth, has not been examined.

The amino-acid mixtures which have so far been prepared are less efficient than native proteins. Block and Bolling consider there are four "semi-indispensable" amino acids: (*a*) arginine and glycine, which are synthesized in the body, but at an insufficient rate to maintain maximum growth; (*b*) cystine and tyrosine, which are indispensable when there are insufficient amounts of methionine and phenylalanine respectively.

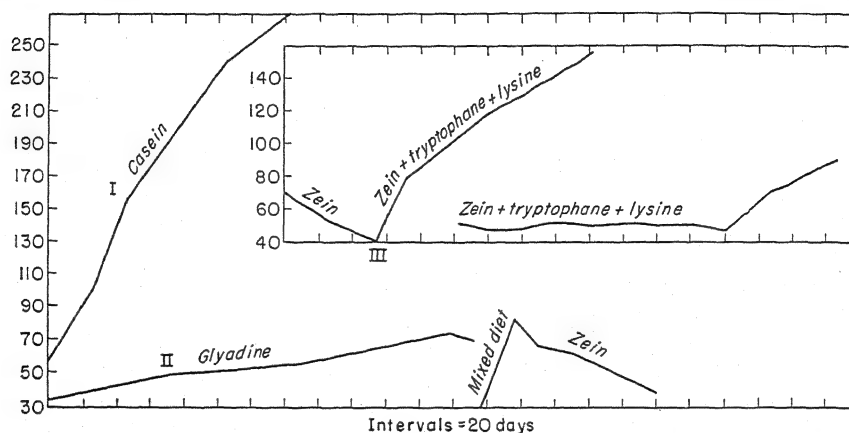



FIG. 204. Growth curves of rats fed on a basic diet to which different proteins were added. Ordinate, weight in grams. Curve I (casein) corresponds to normal growth. (*Osborne, T. B., and L. B. Mendel, J. Biol. Chem., vol. 17, p. 325, 1914.*)

[1] ALBANESE, A. A., *Advances in Protein Chemistry*, 3, 227, 1947.

Table 50. Essential and Nonessential Amino Acids

Essential amino acids				Nonessential amino acids for the rat and man
Name	*For the rat*	*For man, gm. per day*		
		Minimum	*Advisable*	
L-Tryptophane	+	0.25	0.50	L-Alanine
L-Lysine	+	0.80	1.60	L-Aspartic acid
L-Methionine	+	1.10	2.20	L-Citrulline
L-Valine	+	0.80	1.60	L-Glycine
L-Leucine	+	1.10	2.20	L-Glutamic acid
L-Isoleucine	+	0.70	1.40	L-Cystine†
L-Phenylalanine	+	1.10	2.20	L-Hydroxyglutamic acid
L-Threonine	+	0.5	1.0	L-Hydroxyproline
L-Arginine*	+			L-Norleucine
L-Histidine*	+			L-Proline
				L-Serine
				L-Tyrosine‡

* Arginine is synthesized at an insufficient rate to assure normal growth in the rat; it is not essential for man. The dispensability of arginine and histidine, however, is not unanimously accepted (Albanese).

† It can replace methionine in part, but not completely.

‡ It can replace phenylalanine in part, but not completely.

The amount of amino acids in foodstuffs has been determined; there are special tables with the corresponding figures. Also the amino-acid content of normal diets has been established (Table 51).

another protein that contains them. Thus on zein alone growth and weight cannot be maintained, however large the amount given (Fig. 204); neither can normal growth be obtained

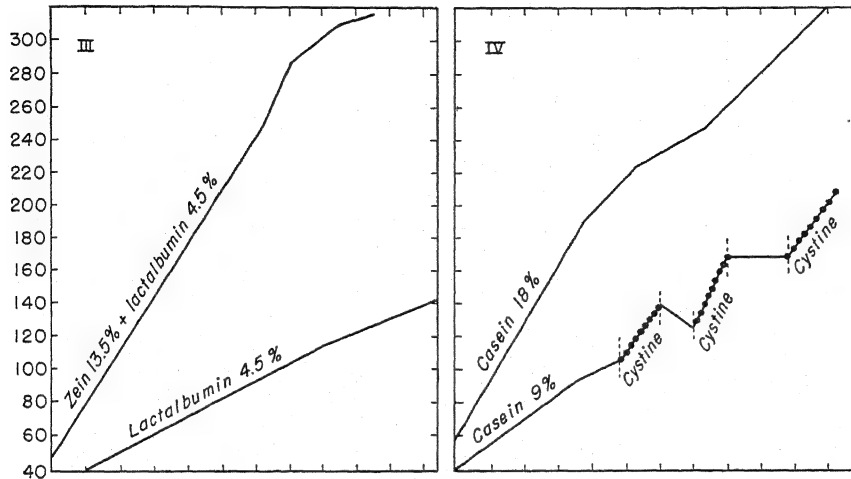



Fig. 205. Growth curves of rats fed on a basic diet to which different proteins were added. Ordinate, weight in grams. In III an insufficient amount (4.5 per cent) of a complete protein—lactalbumin—supplements successfully an incomplete protein—zein. In IV the addition of cystine supplements an insufficient amount (9 per cent) of casein. (*Osborne, T. B., and L. B. Mendel, J. Biol. Chem., vol. 17, p. 325, 1914.*)

when the diet contains less than 8 or 9 per cent lactalbumin as the only protein. Nevertheless, a diet containing 13.5 per cent zein supplemented with 4.5 per cent lactalbumin assures normal growth, because the small amount of lactalbumin supplements the deficiencies of zein (Fig. 205). In corn grain, zein is supplemented by glutelin, a complete protein which contains the amino acids missing in the former.

Supplementary relations between proteins. An incomplete protein, which by itself cannot serve to maintain growth and life, can nevertheless be used in nutrition if the missing amino acids are added to it, or if it is associated with

Table 51. Amino Acids in Grams per Day in Normal Diets of Adults

Amino acids	*American diets*		*French diet, Randoin and Fournier*
	Macy	*Block*	
* Valine	3.2	3.9	4.2
* Leucine	9.6	12.6	8.6
* Isoleucine	3.1	3.7	3.8
* Threonine	3.2	3.6	2.7
* Methionine	3.7	4.1	4.3
* Lysine	4.6	5.2	3.8
* Phenylalanine	4.2	4.7	4.5
* Tryptophane	0.9	1.1	0.9
Histidine	1.6	2.0	2.2
Arginine	4.7	4.7	4.3
Tyrosine	3.9	3.9	3.8

* Essential amino acids.

A diet should contain several proteins, especially if it is a vegetarian diet. If only a small amount of protein can be given, it should be an animal protein of high biologic value. A mixed diet should contain the proteins of eggs, milk, liver or kidney, meat, potato, wheat, oats, corn, or beans (these have been given in decreasing order of nutritive value).

SPECIFIC DYNAMIC ACTION

During the absorption of food there is an increase in the metabolic rate, known as the "specific dynamic action" (SDA) of food. For every 100 cal. produced in basal conditions, 130 cal. is produced if the subject has ingested the equivalent of 100 cal. of meat protein. If the equivalent of 100 cal. of carbohydrate is ingested, 104 cal. will be produced, and 106 cal.

for every 100 cal. of fat. These are only approximate values, which vary from one experiment to another. The extra energy produced by the ingestion of protein cannot be converted into mechanical or other forms of energy; it is lost as heat. Specific dynamic action can be observed only in subjects kept at an environmental temperature of 30°C.; if the animals are at a temperature that stimulates heat production for the maintenance of the body temperature (see Chap. 48), the extra heat of SDA is incorporated into the heat used for body-temperature regulation.

The specific dynamic action of proteins is dependent on the process of deamination that takes place in the liver. Hepatectomy suppresses deamination and the SDA.[1] Six amino acids are responsible for the SDA of proteins. These are glycine, alanine, leucine, glutamic acid, tyrosine, and phenylalanine; the last of these is the most active (Rapport and Beard). The intensity of SDA produced by a protein depends on the percentage of these amino acids it contains.

SDA of proteins does not seem to depend on the utilization of the deaminated residue. Thus in a dog treated with phlorhizin, proteins and amino acid produce an SDA, although all the extra glucose formed is excreted in the urine (Lusk). It is not due to digestive activity, because it is not produced by stimulation of the secretions or motility of the digestive tract. Several hypotheses have been put forward to explain the mechanism of SDA. It is generally admitted to be due to the six amino acids mentioned, to the NH_2 group (Grafe), and to certain products of the intermediary metabolism of those acids.[2]

Lundsgaard[3] observed that the oxygen consumption of the perfused liver of a cat increased at the same rate as urea formation when amino acids or ammonium salts were added to the perfusion fluid. He concluded that the increase in O_2 consumption was due to mobilization of energy for the formation of urea.

SDA of fats has been attributed to fat plethora, and it lasts 10 to 12 hr. That of carbohydrate is also attributed to plethora, which increases glucose consumption and its conversion into glycogen.

[1] MANN, F. C., C. M. WILHELMJ, and J. L. BOLLMAN, *Am. J. Physiol.*, **81**, 496, 1927.

[2] LUSK, G., *J. Nutrition*, **3**, 519, 1931; WILHELMJ, C. M., *Physiol. Rev.*, **15**, 202, 1935; SCHAEFFER, G., and E. LE BRETON, "L'action dynamique spécifique des protides," Hermann & Cie, Paris, 1938; BORSOOK, H., *Biol. Rev.*, **11**, 147, 1936.

[3] LUNDSGAARD, E., *Acta physiol. Scandinav.*, **4**, 330, 1942.

SDA is diminished in thyroid and hypophyseal insufficiencies in certain species; *e.g.*, it is diminished in hypophyseal insufficiency in man and rats, but not in dogs. This seems to be due not to a direct effect of hormonal insufficiencies on metabolism but to retarded intestinal absorption.

PROTEIN STORAGE

Carbohydrate and fat are stored in the organism, but there is no satisfactory demonstration of protein storage. The so-called "protein stores" are constituted mainly by plasma proteins and by labile protein in the cytoplasm of the cells of muscle, liver, lymphoid tissues, etc. In an emergency, such as fasting, protein catabolism increases and many cells release protein, but this is recovered on return to normal conditions. Thus, in conditions of stress adrenal hormones are secreted in excess and lympholysis, *i.e.*, disintegration of lymphocytes and lymphoid cells, takes place. Thyroid hormone seems to play a similar part in the rapid mobilization of muscle protein. Increase of body proteins observed in growth, or when muscles hypertrophy owing to exercise, or in pregnancy and lactation, or after a prolonged fast, or in convalescence cannot be considered as storage of proteins, but as an increase in body mass.

THE FATE OF AMINO ACIDS

After absorption amino acids undergo one of the following processes: (*a*) *deamination*, *i.e.*, separation of the NH_2 with the formation of nitrogenous catabolic end products such as urea and creatinine in mammals and uric acid in birds; (*b*) *transamination*, *i.e.*, passage of NH_2 from one amino acid to another compound; (*c*) *utilization of the deaminated residue*, which may be converted into glucose, acetoacetic acid, or other substances and subsequently metabolized; (*d*) *utilization in protein synthesis* (plasma and tissue proteins, hemoglobin, etc.).

The following are some of the methods that have been used in the study of amino-acid metabolism. (*a*) observation of growth, nitrogen balance, and general health of animals fed on diets in which certain amino acids are absent or have been added; (*b*) determination in blood and urine of products of amino-acid catabolism; (*c*) perfusion of isolated organs; (*d*) transformations undergone by amino acids in the presence of surviving slices of tissues (Krebs); (*e*) permanent fistulization of blood vessels

(London's angiostomy) by means of silver cannulas, so that blood can be obtained from the arteries or veins (renal vessels, portal and suprahepatic veins, etc.) of healthy unanesthetized animals, in order to determine the concentration (not the rate) at which different substances enter or leave the organ in question; (*f*) administration of amino acids "labeled" by isotopes, such as N^{15} and deuterium in the NH_2 group or C^{14} in the fatty-acid chain, which are then followed throughout the organism; (*g*) observation of certain metabolic anomalies and diseases, such as cystinuria; (*h*) observation of the metabolic processes of yeasts, bacteria, and other simple organisms, which has given valuable information on amino-acid metabolism; (*i*) paper chromatography and utilization by microorganisms are frequently used for the isolation and exact quantitative determination of amino acids.

Amino acids → Deamination → Non-N residue → Carbohydrate; Fat
Deamination → N residue → Transamination; Urea
Amino acids → Protein synthesis → Liver protein → Plasma protein → Tissue protein
Protein synthesis → Tissue protein
Liver protein → Hemoglobin

Temporary storage in the tissues. Amino acids are rapidly taken up from the blood by the tissues—mainly by the liver, where they are rapidly deaminized, the NH_2 group being converted into urea. In hepatectomized dogs, amino acids are not metabolized.

Protein synthesis. In rats fed on amino acids as the only source of nitrogen, their growth shows that the structural proteins of the tissues are synthesized from amino acids. In some cases this synthesis takes place at a very rapid rate; *e.g.*, after extirpation of 70 per cent of the liver in rats, the liver tissue is almost completely regenerated in 9 to 12 days. Another demonstration of the rate of protein synthesis has been obtained by administrating amino acids labeled with deuterium to rats. In 3 days, 10 per cent of the protein in the liver and 2.5 per cent of that in the muscles was found to be of recent formation.[1]

Certain anterior hypophyseal extracts provoke giantism and therefore an increase in body protein formation, similar to that observed in some cases of hyperfunction of the anterior hypophysis (giantism, acromegaly; see Chap. 52, The Hypophysis). Following the injection of these extracts, nonprotein nitrogen and amino acids in plasma decrease, as if the latter were fixed by the tissues at a higher rate than the normal.

The liver is the main site of production of plasma proteins. This must proceed at a rapid rate, because patients with kidney disease may lose 10 to 20 gm. of serum albumin daily for many months, without a decrease in plasma-protein concentration. Half the plasma proteins can be removed from the blood by plasmapheresis,[2] and the effects of different factors on their regeneration can thus be studied. Plasma proteins are rapidly regenerated and the normal concentration can be restored in 2 to 7 days. The administration of blood plasma, casein or plasma hydrolysate, or mixtures of amino acids, by mouth or intravenously, increases the rate of plasma-protein regeneration. Normal regeneration of plasma proteins is dependent on the diet. If an insufficient amount of protein is given there is hypoproteinemia, which, if sufficiently pronounced, causes generalized edema. Severe hepatic disturbances, caused experimentally or by disease, retard the formation and regenera-

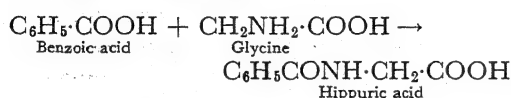
[1] Ussing, H. H., *Nature*, **142**, 399, 1938.

[2] Blood is drawn and centrifuged, the plasma is discarded, and the erythrocytes are then suspended in saline and reinjected. This is repeated several times.

tion of plasma proteins. Regeneration of hemoglobin has priority over regeneration of plasma proteins.

Amino-acid synthesis. At one time amino-acid catabolism alone was studied, but later amino-acid synthesis was well demonstrated[1] for the nonessential amino acids; arginin is also synthesized at a slow rate. Synthesis of glycine is demonstrated by the following facts:

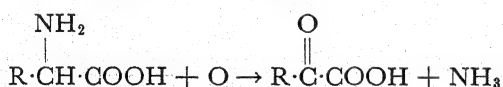
1. Milk proteins contain 0.1 to 0.3 per cent glycine; the proteins of sucklings contain 2.5 per cent glycine;
2. If benzoic acid is given to an animal it combines with glycine and is eliminated in the urine as hippuric acid:

$$\underset{\text{Benzoic acid}}{C_6H_5\cdot COOH} + \underset{\text{Glycine}}{CH_2NH_2\cdot COOH} \rightarrow \underset{\text{Hippuric acid}}{C_6H_5CONH\cdot CH_2\cdot COOH}$$

After a sufficiently long period of treatment more glycine will have been excreted than that contained in the whole organism at the beginning of the experiment plus that ingested while the experiment lasted.

Synthesis of glutamic, aspartic, and other amino acids has also been demonstrated. The formation of alanine has been observed in a surviving liver perfused with fluids containing ketone derivatives, ammonium salts, ammonium pyruvate, or ammonium lactate (Embden).

Deamination and transamination. The greater part of amino acids absorbed undergo deamination. The NH_2 group is split off and converted into urea; a small amount can be used in the formation of other amino acids. The deaminated residue is an α-ketoacid. It constitutes 90 to 95 per cent of the total mass and energy in the original amino acid.

Deamination can be produced by reduction, hydrolysis, or oxidation. In the organism it is mostly performed by oxidation, as follows:[2]

$$R\cdot\underset{|}{\overset{NH_2}{CH}}\cdot COOH + O \rightarrow R\cdot\overset{O}{\overset{\|}{C}}\cdot COOH + NH_3$$

Surviving tissue slices of liver or kidney also deaminize amino acids, but other tissues do not. In the dog the liver is the only important site of deamination and urea formation; neither of these processes occurs after hepatectomy in this species (Mann *et al.*).

Transamination consists in the transfer of NH_2 from an amino acid to another chemical compound, *e.g.*, a ketoacid, with the formation of another amino acid.[1] The enzymes that catalyze these reactions are known as transaminases, and pyridoxal phosphate acts as coenzyme. Schoenheimer and his associates[2] have shown that these processes are constantly taking place in the organism. A large proportion of absorbed N is incorporated into the tissues and there is a continuous dynamic exchange of nitrogen in the tissues. The idea that there are two different protein metabolisms—an exogenous protein metabolism, in which ingested protein is concerned, independent of the endogenous protein metabolism of tissue protein—can no longer be accepted. Protein is continuously being formed and decomposed in the cells. Moreover, nitrogen is transferred from one amino acid to another; thus, if leucine or glycine labeled with N^{15} is given, the isotope will be found in glutamic and aspartic acids, and in small amounts in valine but not in lysine.

Administration of N^{15} in ammonium citrate has shown that N^{15} is also transferred from this substance to amino acids. There is also proof of the formation and disintegration of amino-acid chains. Glutamic acid is the most active of the amino acids in this process.

Fate of the deaminized residue. After deamination the ketoacid residue is finally oxidized into CO_2 and water. In animals treated with phlorhizin it has been possible to demonstrate that this residue follows one of three paths. The so-called "glycogenic" amino acids form extra glucose and are excreted in the urine. They all have two to five carbon atoms (with the exception of arginine, which has six), and with the exception of lysine, they all have a straight chain; only one (proline) is a cyclical amino acid. The following are glycogenic amino acids: glycine; alanine; serine; threonine; cystine; aspartic, glutamic, and hydroxyglutamic acids; arginine (only three of its C atoms);

[1] KNOOP, F., *Ztschr. f. physiol. Chem.*, **67**, 489, 1910; **89**, 151, 1914; *Klin. Wchnschr.*, **17**, 1309, 1938.

[2] KREBS, H. A., *Ztschr. f. physiol. Chem.*, **217**, 191, 1933; *Biochem. J.*, **26**, 1620, 1935.

[1] BRAUNSTEIN, A. E., *Advances in Protein Chemistry*, **3**, 1, 1947; GUNSALUS, I. C., *Federation Proc.*, **9**, 556, 1950; CAMMARATA, P. S., and P. C. COHEN, *J. Biol. Chem.*, **187**, 439, 1950.

[2] SCHOENHEIMER, R., S. RATNER, and D. RITTENBERG, *J. Biol. Chem.*, **127**, 333, 1939; **130**, 703, 1939.

valine (only three of its C atoms); ornithine; and proline.

Ketogenic amino acids form acetoacetic acid. The following belong to this group: leucine, isoleucine, tyrosine, and phenylalanine.

Tryptophane, histidine, methionine, and lysine are neither glycogenic nor ketogenic.

THE END PRODUCTS OF PROTEIN CATABOLISM

Urinary excretion. The end products of protein catabolism are excreted in the urine. In mammals and other ureothelic animals, urea is the main end product; in birds and reptiles it is uric acid, and they are known as uricothelic animals.

The amount of total nitrogen, urea, and total sulfur excreted is in direct proportion to the amount of protein in the diet. The amount of urine excreted is in direct proportion to the amount of nitrogenous substances excreted. Folin supposed that these substances were the end products of ingested protein. This theory can no longer be accepted, because at least one-half of the nitrogen absorbed with food passes into the tissues and is not catabolized immediately. Urea excreted by fasting or fed animals is not formed exclusively from ingested amino acids; part, at least, derives from tissue proteins.

Protein seems to be actively catabolized. Thus an amount of N equivalent to 50 per cent of that which has been ingested in a meal is excreted by the dog in 4 to 5 hr.; by 12 to 16 hr. an amount equivalent to all the N ingested has been eliminated.

ingested. The normal range of urea concentration in blood is from 18 to 35 mg. per 100 cc. (equivalent to 8 to 15 mg. N). A concentration of 50 mg. per cent is definitely abnormal. An increase in blood urea is observed when it is not normally eliminated by the kidney, when there is insufficient elimination of water, or when it is formed in excess. The highest concentration that urea can reach in human urine is 55 gm. per liter; therefore if water is excreted in very low quantities, urea will be retained in the body.

Formation of urea by the liver has been demonstrated in many ways: (*a*) the liver has been perfused with blood to which ammonium salts or amino acids have been added; (*b*) urea is formed by tissue slices if amino acids or sodium salts are added; (*c*) deamination of amino acids and urea formation cease completely after hepatectomy in dogs:[1] (*d*) in patients with severe forms of hepatic disease, in which there is extensive destruction of liver cells, urea excretion diminishes to a minimum, and apparently little or no urea is formed.

The principal mechanism of urea formation seems to be that discovered by Krebs and Heinseleit[2] in which arginine is the precursor of urea. The NH_2 groups that have been split off from amino acids are converted into urea in three steps: (*a*) ornithine combines with ammonia and forms citrulline and water; (*b*) one molecule of citrulline combines with one molecule of ammonia and forms arginine and water; (*c*) arginine is hydrolyzed by arginase into ornithine and urea.

$$\underset{\text{Ornithine}}{COOH{\cdot}CHNH_2{\cdot}CH_2{\cdot}CH_2{\cdot}CHNH_2} + NH_3 + CO_2 \rightarrow \underset{\text{Citrulline}}{COOH{\cdot}CHNH_2{\cdot}CH_2{\cdot}CH_2{\cdot}CHNH{\cdot}CONH_2}$$

$$\downarrow + NH_3 \text{ from aspartic acid}$$

$$\underset{\text{Urea}}{CO(NH_2)_2} \leftarrow H_2O + \underset{\text{Arginase}}{} \underset{\text{Arginine}}{COOH{\cdot}CHNH_2{\cdot}CH_2{\cdot}CH_2{\cdot}CHNH{\cdot}C(NH)NH_2} \quad (\text{Ornithine} \uparrow)$$

Urea. A subject on a mixed diet excretes from 9 to 13 gm. of nitrogen daily. Most of this N is in the form of urea. If the subject is fed a mixed diet, 87 per cent of urinary N is in the form of urea; 95 per cent if a high-protein diet, and 60 per cent if a low-protein diet is ingested. Subjects on a mixed diet excrete 20 to 30 gm. of urea daily.

Urea is formed from the NH_2 group split off by deamination from amino acids; it is also formed from ammonium salts when these are

The process is not as simple as this; seven stages, with the intervention of several enzymes, are already known.[3] Aspartic acid is the NH_3 donor for the conversion of citrulline to arginine; other amino acids

[1] BOLLMAN, G. L., F. C. MANN, and T. B. MAGATH, *Am. J. Physiol.*, **69**, 371, 1924; *Ergebn. d. Physiol.*, **23**, 212, 1924.

[2] KREBS, H. A., and K. Z. HEINSELEIT, *Ztschr. f. physiol. Chem.*, **210**, 33, 1932.

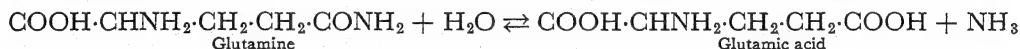
[3] RATHNER, S., and S. GRISOLIA, Urea Formation A Symposium on Phosphorus Metabolism," Johns Hopkins Press, Baltimore, 1951.

can transfer it to aspartic acid by transamination. Carbamil glutamate gives CO_2 and NH_3 for the conversion of ornithine into citrulline. ATP (adenosine-triphosphate) takes part in all these reactions.

Ammonia. A subject on a mixed diet excretes in the urine from 0.4 to 1 gm. ammonia (2 to 5 per cent of total urinary N) every 24 hr.[1] The amount depends on the acid-base equilibrium of the organism. Thus it increases (*a*) if acids, such as HCl, which are not metabolized, are ingested; (*b*) if acids are formed in the organism, as when a protein diet is given, in acidosis, muscular exercise, pregnancy, or fever; (*c*) if there is diabetic acidosis. In this last case, there is a very great increase because β-hydroxybutyric and acetoacetic acids are excreted as ammonium salts. Urinary ammonia is diminished by (*a*) ingestion of alkali; (*b*) a vegetarian diet, which has more basic than acid end products.

In shed blood small quantities of free ammonia are found, but it is doubtful if there is any in circulating blood. Immediately after extraction 0.1 mg. per cent (Parnas and Heller) or 0.5 mg. per cent (Krebs) has been reported; if the blood is collected in an atmosphere of CO_2, only 0.004 mg. per cent is found (Conway).

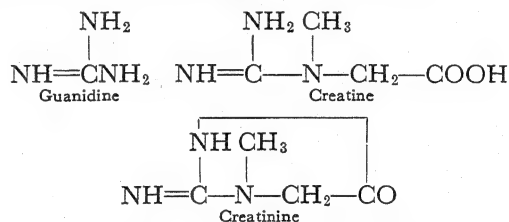
Nash and Benedict[2] demonstrated that ammonia is formed in the kidney, and that there is more ammonia in the blood of the renal vein than in renal arterial blood. Renal production of ammonia prevents the loss of base in the urine; sodium and potassium salts are converted into ammonium salts in the kidney, the acid is thus eliminated, and the base is retained in the organism. Ammonia is formed from glutamine (the amide of glutamic acid) and not from urea, which is eliminated as such.[3]

$$\underset{\text{Glutamine}}{COOH{\cdot}CHNH_2{\cdot}CH_2{\cdot}CH_2{\cdot}CONH_2} + H_2O \rightleftarrows \underset{\text{Glutamic acid}}{COOH{\cdot}CHNH_2{\cdot}CH_2{\cdot}CH_2{\cdot}COOH} + NH_3$$

If ammonia is given to an animal it is converted into urea (mammals) or into uric acid (birds). In man and other mammals, when ammonium chloride is given, NH_4 is converted into urea and excreted; Cl^- is retained in the body, where it combines with fixed base (bicarbonate) and causes a decrease in the alkali reserve and acidosis.

Hippuric acid. There is only a small amount of this acid in human urine (0.7 to 1 gm. in 24 hr., *i.e.*, 3 to 5 per cent of total urinary N). It increases in subjects on a vegetarian diet (2 gm. or more). It is found in large quantities in the urine of herbivorous animals, which excrete from 50 to 150 gm. daily. It is formed by the combination of glycine and benzoic acid, a process that takes place in the kidney in some species (Bunge and Schmiedeberg), but mainly in the liver in most animals. Formation of hippuric acid after ingestion of benzoate is used as a test of hepatic function in man (Quick).

Creatine and creatinine. Creatine is methylguanidinacetic acid. Creatinine (methylglycocyamidine) is the anhydride of creatine.

$$\underset{\text{Guanidine}}{NH{=}C(NH_2)NH_2} \qquad \underset{\text{Creatine}}{NH{=}C(NH_2){-}N(CH_3){-}CH_2{-}COOH}$$

$$\underset{\text{Creatinine}}{NH{=}C{-}N(CH_3){-}CH_2{-}CO \text{ (ring closed through NH between C and CO)}}$$

Creatine is a constituent of the tissues and creatinine is excreted in the urine. There is 90 to 120 gm. of creatine in the human body; 98 per cent of this is in the muscles, 1.5 per cent in nerve tissue, and the rest in the testicle and other tissues. Muscle contains 350 to 400 mg. of creatine per 100 gm.; 80 per cent of this is in the form of phosphocreatine, a substance of great importance in the energetics of muscular contraction. Blood really contains very little creatine, but the methods currently used for its determination give artificially high results. Creatine is seldom found in urine.

There is very little creatinine in the tissues, but it is regularly found in urine. It is an end product of metabolism that the organism cannot utilize. Thus if creatinine labeled with N^{15} is administered to an animal, all of it will be found in the urine. If, on the other hand, creatine labeled with N^{15} is administered it is not found in the urine, unless very large amounts are given. It is retained in the tissues and is later excreted

[1] If urine is left standing at room temperature without an antiseptic, ammonia is formed from urea by bacterial action.

[2] Nash, T. P., *J. Biol. Chem.*, **48,** 663, 1921; **51,** 183, 1922; Nash, T. P., and S. R. Benedict, *J. Biol. Chem.*, **69,** 381, 1926; **82,** 673, 1929.

[3] Van Slyke, D. D., R. A. Phillips, P. B. Hamilton, R. M. Archibald, P. M. Futcher, and A. Hiller, *J. Biol. Chem.*, **150,** 481, 1943.

as creatinine. The administration of protein or amino acids increases formation and excretion of creatine and creatinine in the rat. This fact has led Beard to suppose that all the amino acids can produce creatine, but Rose believes this is due to stimulation of creatine metabolism and that amino acids are not necessarily creatine precursors.

More recent studies[1] have led to the conclusion that creatine is formed by the interaction of three substances: glycine, arginine, and methionine. Glycine is absorbed with the food or formed in the course of metabolism. Arginine is the source of the guanidine group. Methionine, choline, or other substances act as methyl donors. The process is as follows:

$$\underset{\text{Arginine}}{R\cdot C(NH_2)=NH} + \underset{\text{Glycine}}{CH_2(NH_2)-COOH} \rightarrow \underset{\text{Guanidinacetic acid}}{NH=C(NH_2)-NH-CH_2-COOH} + \underset{\text{from Methionine}}{CH_3} \rightarrow \underset{\text{Creatine}}{NH=C(NH_2)-N(CH_3)-CH_2-COOH}$$

Creatinine is, with the exception of urea, the most abundant of the nitrogenous substances in urine. An adult man excretes 1 to 1.5 gm. daily (3 to 6 per cent of total urinary N). Folin considered creatinine excretion to be constant, and established a "creatinine coefficient," *i.e.*, the amount excreted per kilogram per hour; normally he found this to be 20 to 26 in man, 14 to 22 in women, and 30 in athletes. Apparently there is some relation between creatinine excretion and muscular development. Creatinine arises from creatine (phosphocreatine) in muscle, which is maintained at a constant level and is metabolized at a constant rate, even during fasting. The conversion of phosphocreatine into creatinine, with liberation of phosphorus, has been demonstrated recently. Probably this is the path usually followed in the formation of creatinine by animal tissues.

The only significant precursors of creatinine are guanidinacetic acid and glycine. Guanidinacetic acid is methylated and creatine formed only in liver tissue. In the disease known as "pseudo-hypertrophic progressive muscular dystrophia" usually there is creatinuria, which increases if glycine is given; in some cases glycine administration has been followed by improvement of the condition.[2]

Creatine is not a normal constituent of adult human urine. Creatinuria may be observed in the following circumstances: (*a*) commonly in children; (*b*) sometimes in women, regularly during pregnancy, puerperium (even if hysterectomy has been performed), and lactation; (*c*) in complete fasting, carbohydrate fasting, and diabetes; (*d*) in thyrotoxicosis and in fever; (*e*) in poliomyelitis, muscular dystrophia, myasthenia, and certain mental diseases. According to Beard, creatinuria is related to disturbances in phosphorus metabolism.

Blood serum has small amounts of creatine (0.42 mg. per cent) and creatinine (1.07 mg. per cent). An increase of these substances in the serum is a sign of nitrogen retention in renal insufficiency.

Creatinine excretion increases in inanition, fever, thyrotoxicosis, and diabetes. Creatinine in the blood plasma is filtered through the renal glomerulus, and in man a small amount is added in the renal tubes. In renal diseases with disturbance of nitrogen excretion, blood creatinine increases because of faulty elimination. Creatinine concentration in blood can therefore be taken as a sign of the state of nonprotein nitrogen excretion.

$$\text{Creatine} + PO_4^{-3} \rightarrow \underset{\text{Phosphocreatine}}{HN=C(NHPO(OH)_2)-N(CH_3)-CH_2\cdot COOH} \rightleftarrows \underset{\text{Creatinine}}{HN=C\langle NH-CO-CH_2-N(CH_3)\rangle}$$

The metabolism of sulfur. Almost all the proteins contain sulfur combined in cystine and methionine. Sulfur is particularly abundant in the protein of the hair and nails. It is also found in some glucoproteins, *e.g.*, in mucoitinsulphuric acid in mucin and in chondroitinsulfuric acid in

[1] Bloch, K., and R. Schoenheimer, *J. Biol. Chem.*, **138**, 155, 1941.

[2] Brand, E., *et al.*, *Am. J. Physiol.*, **90**, 296, 1929.

cartilage. Cystine gives rise to cysteine, glutathione (a tripeptide formed by glutamic acid, cysteine, and glycine), and taurine, which combines with cholalic acid to form the taurocholic acid in bile.

Cystine and methionine are completely oxidized into sulfates by the tissues. Methionine is one of the essential amino acids that must be included in the diet, although man needs relatively less than other species. In some animals

$$CH_3-S-(CH_2)_2-CHNH_2-COOH$$
Methionine

$$HOOC-CHNH_2-CH_2-S-S-CH_2-CHNH_2-COOH$$
Cystine

$$SH-CH_2-CHNH_2-COOH$$
Cysteine

$$SO_2OH-CH_2-CH_2NH_2$$
Taurine

it increases the retention of nitrogen in the organism, but not in man. It is converted into cystine by the tissues. Methionine acts as a methyl donor[1] in the formation of choline, creatine, and other substances. Methionine in the diet protects the liver from necrosis and cirrhosis produced by low-protein–high-fat diets.

Sulfur is also a component of vitamin B_1 (thiamine) and some of the hormones (insulin, anterior hypophyseal hormones).

Sulfur is found in three forms in the urine: (*a*) inorganic sulfur, (*b*) ethereal sulfate; (*c*) organic, or neutral, sulfur.

In the course of metabolism the greater part of cystine and methionine is oxidized, and sulfates of K, Na, Ca, and NH_3 are formed; this makes up 80 to 90 per cent of total urinary sulfur.

Ethereal sulfates (5 to 8 per cent of total urinary sulfur) are compounds of sulfuric acid and phenol or cresol derived from phenylalanine or tyrosine, and indoxyl or skatoxyl derived from tryptophane. These substances are formed by the action of bacteria in the intestine; part of these bacterial products are absorbed and combined with sulfuric acid. Sulfoconjugation of phenols is performed in many tissues, *e.g.*, the intestine, liver, etc., but it is still actively carried out even in hepatectomized animals. Sulfur compounds of phenol are less toxic than phenol, and for this reason sulfoconjugation is considered a disintoxicating or protective mechanism. The elimination of ethereal sulfates in urine (urinary indican) indicates the rate of intestinal putrefaction. Intestinal antiseptics diminish or even suppress urinary indican.

Neutral sulfur (4 to 6 per cent of total urinary sulfur) consists of cystine, urochrome sulfocyanide, and other organic compounds of sulfur.

Inorganic urinary sulfur is, together with urea, an index of protein catabolism. The excretion of sulfate and of urea varies simultaneously and in the same direction, except for a few transitory discrepancies.

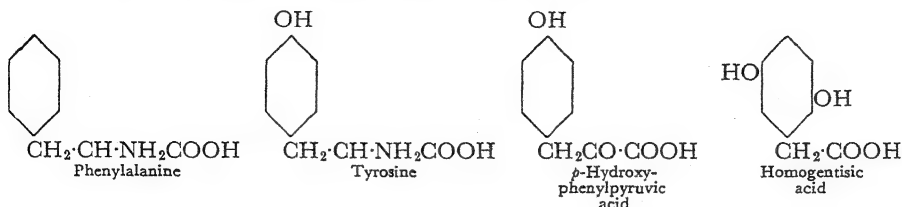
Certain individuals excrete cystine in the urine. This rare metabolic anomaly is called cystinuria. Not only is there a disturbance in cystine metabolism, but other amino acids (lysine, arginine, ornithine) also increase in the blood and urine. Cystine is almost insoluble in water; therefore it crystallizes in the urine (urinary "sand") or forms stones in the urinary tract. Crystals of cystine may also be found in the bone marrow and other tissues.

CATABOLISM OF CYCLIC AMINO ACIDS

Phenylalanine and tyrosine. These amino acids are constituents of a large number of proteins. Phenylalanine is one of the essential amino acids that must be included in the diet. The administration of phenylalanine labeled with deuterium shows that it is converted into tyrosine, but tyrosine is not converted into phenylalanine. Oxidative deamination of phenylalanine and tyrosine gives *p*-hydroxyphenylpyruvic acid; this process has been obtained *in vitro* with slices of surviving tissues. A single case has been reported by Medes (1932) in which a patient eliminated *p*-hydroxyphenylpyruvic acid in the urine spontaneously or following the administration of phenylalanine or tyrosine. The disturbance has been called "tyrosinosis." In normal subjects the benzene ring of these amino acids is broken, and acetoacetic

[1] Du Vigneaud, V., "Interrelationships between Choline and Other Methylated Compounds, *Biol. Symposia*, 5, 234, 1941.

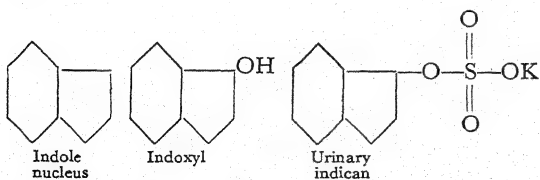
acid is formed; this has been demonstrated by experiments in phlorhizinized dogs.

In certain subjects alkaptone (homogentisic or 2.5-dihydroxyphenylacetic acid) is eliminated in the urine, which becomes black when left standing in contact with air or when alkali is added. The urine reduces Fehling's and Benedict's reagent and gives a violet color with ferric chloride. Alkaptonuria increases after the administration of phenylalanine or tyrosine. Deposition of alkaptone in tissues (sclerotic, cartilage) produces a gray or brown pigment (ochronosis). This metabolic anomaly is inherited as a mendelian recessive character.

$CH_2{\cdot}CH{\cdot}NH_2COOH$ Phenylalanine; OH / $CH_2{\cdot}CH{\cdot}NH_2COOH$ Tyrosine; OH / $CH_2CO{\cdot}COOH$ *p*-Hydroxyphenylpyruvic acid; HO, OH / $CH_2{\cdot}COOH$ Homogentisic acid

Phenylalanine is the precursor of tyrosine, of adrenaline produced in the adrenal medulla and of diiodotyrosine and thyroxine produced by the thyroid. Tyrosine and adrenaline can form black or red pigments.

Phenylketonuria. Phenylpyruvic acid has been found in the urine of mentally deficient patients (0.5 per cent of the inmates in an asylum for mental deficients) who suffered from a mental condition which has been called phenylpyruvic imbecility (Folling, 1934). Phenylpyruvic acid is formed in the kidney from phenylalanine which is not converted into tyrosine.

Tryptophane. This is one of the essential amino acids. It can give rise to a certain amount of the pellagra-preventing (P-P) vitamin. Administration of tryptophane causes the formation of kinurenic acid in some species, but not in man. Intestinal bacteria form indole and skatole from tryptophane, which are partially reabsorbed. Indole is converted into indoxyl by the tissues, mainly in the liver, and then into potassium indoxylsulfate (urinary indican), which is found in the blood and tissue fluids and is eliminated by the kidney. It is an index of intestinal putrefaction.

Indole nucleus; —OH Indoxyl; —O—S(=O)(=O)—OK Urinary indican

If there is a disturbance in renal excretion its concentration in blood will be found to be persistently high, and it will be eliminated at a slow rate after injection. Retention of indican usually occurs simultaneously with that of nonprotein nitrogenous substances.[1]

Pyrollic amino acids. Proline and oxyproline are completely disintegrated in the organism following the path of the tricarboxylic cycle. Three of the carbon atoms can be utilized for the formation of glucose.

Histidine. This is imidazol-α-aminopropionic acid, which is completely oxidized in the body. A small amount may be found in normal urine, and larger quantities in the urine of pregnant women. In shock, in allergic reactions, and by the action of bacteria, histidine can give rise to histamine (β-imidazolethylamine). This substance has a potent effect: it can produce shock. It contracts smooth muscles, dilates the capillaries, and increases the passage of fluid from the blood to the tissues. In all tissues, except the lung, an enzyme that destroys histamine has been found; it is called histaminase. Histidine is considered a precursor of purine bases in the body.

ENDOCRINE REGULATION OF PROTEIN METABOLISM

Certain hormones, in adequate amounts, are indispensable for normal growth and therefore for the formation of tissue proteins. In thyroid and hypophyseal insufficiencies, growth is retarded and less nitrogen is retained for protein formation than in normal subjects. Protein catabolism is also retarded and in complete fasting or in protein fasting less nitrogen is eliminated in the urine. Anterior hypophyseal or thyroid treatment restores the normal condition in the respective insufficiencies.

Testicular hormones are also important in the stimulation of protein anabolism. The muscular development of castrates is less than that of normal individuals, and in many cases testosterone causes retention of nitrogen, even in normal subjects.

[1] An examination of substances that give rise to indoxyl or indoleacetic acid has been made by Stoppani (*J. Biol. Chem.*, **157**, 1, 1945).

In cases of hyperfunction of the anterior hypophysis (giantism, acromegaly) there is an increase in the development of the body, due to an increase in protein anabolism (see Chap. 52, The Hypophysis).

Giantism can be produced in rats and dogs by the administration of anterior hypophyseal extract. After each injection, nonprotein nitrogen in the blood and urinary excretion of nitrogen diminish. This effect is apparently due to the taking up of nitrogen by the tissues for the synthesis of protein.

In the first stages of certain cases of corticoadrenal hyperfunction, there is a precocious and abnormally great development of the muscles, even in women and children.

Small doses of thyroid can transitorily improve growth in some cases, but hyperthyroidism, especially in severe cases, causes an increase in protein catabolism, the minimum protein requirement rises, and creatinuria may be observed. The adrenals play an important part in the increase in protein catabolism observed in diabetes, anoxia, and other conditions. Some of the anterior hypophyseal and corticoadrenal hormones increase the formation of glucose from protein when the subject is fasting or when there is already extra glucose formation (diabetes). Adrenalectomy or hypophysectomy, on the contrary, diminishes nitrogen excretion, hyperglycemia and glycosuria in diabetic animals.

THE METABOLISM OF NUCLEOPROTEINS

Nucleoproteins are conjugated proteins; hydrolysis splits them into a nuclein and a protein. Nucleins are formed by a nucleic acid and a basic protein (histone or protamine). The best known of the nucleic acids are thymus nucleic acid and yeast nucleic acid. At one time these were considered as typical of animal and plant organisms respectively, but later thymus nucleic acid was found in plants, and yeast (or ribonucleic) acid was found in animals.

Nucleic acids were given the name because they were first isolated from nucleoproteins obtained from cell nuclei.[1] Later, nucleic acids were found in the cytoplasm.

The nucleic acid in chromatin has desoxyribose in its molecule; it is known as "desoxyribonucleic" or "chromonucleic" acid. In the cytoplasm and nucleoli of animal cells there is nucleic acid with ribose in its molecule, known as "ribonucleic" or "plasmonucleic" acid.

Desoxyribonucleic acid gives a positive Feulgen test, *i.e.*, fuchsin bleached by sulfurous acid recovers its red color. The nucleus and chromosomes also give a positive Feulgen test. Chromatin owes to nucleic acid its affinity for basic stains. Nucleic acids can be identified in sections of tissue by their absorption spectra (Caspersson).

Nucleic acids are of considerable biologic importance. They form part of the specific protein of viruses and of chromosomes. The latter constitute the physical basis of heredity and play a leading part in cell division and in the metabolic and secretory activities of the cells. Nucleic acids are apparently of fundamental importance for the formation of the specific substances of the organism.

Table 52. Composition of Nucleic Acid

Constituents	*Thymus and nuclear chromatin; desoxyribonucleic acid*	*Yeast and cytoplasm; ribonucleic acid*
Phosphoric acid...	Phosphoric acid	Phosphoric acid
Sugar...........	Desoxyribose	*d*-Ribose
Purine bases......	Adenine and guanine	Adenine and guanine
Pyrimidine bases...	Cytosine and thymine	Cytosine and uracil

A nucleic acid unit is made up of four nucleotides. Each nucleotide is formed by (*a*) phosphoric acid; (*b*) a pentose; (*c*) a nitrogenous base (Table 52). Two of the nitrogenous groups are purine bases, adenine and guanine, and two are pyrimidine bases, thymine and cytosine. According to Levene, desoxyribonucleic acid (chromonucleic) or thymus nucleic acid is made up of the following four nucleotides:

Adenine-desoxyribose-phosphoric acid
= adenine-nucleotide or adenylic acid
Thymine-desoxyribose-phosphoric acid
= thymine-nucleotide
Cytosine-desoxyribose-phosphoric acid
= cytosine-nucleotide
Guanine-desoxyribose-phosphoric acid
= guanine-nucleotide or guanylic acid

[1] Nucleic acid was discovered by Miescher (1871) and purines by Kossel (1879–1881). The structure of purines was first described by Emil Fischer (1907) and the constitution of tetranucleotides by Levene (1909–1931).

```
N1====6CH
|      |
HC2   5C—7NH
||     ||    \
||     ||     8CH
||     ||    //
N3———4C———9N
```

Purine

```
N===C—NH2
|   |
HC  C—NH
||  ||   \
||  ||    CH
||  ||   //
N———C———N
```

Adenine
6-amino-purine

```
     HN———CO
     |    |
NH2—C     C—NH
    ||    ||   \
    ||    ||    HC
    ||    ||   //
    N———C———N
```

Guanine
2-amino-6-oxypurine

```
HN———CO
|    |
HC   C—NH
||   ||   \
||   ||    CH
||   ||   //
N————C———N
```

Hypoxanthine
6-oxypurine

```
HN———CO
|    |
CO   C—NH
|    ||   \
|    ||    CH
|    ||   //
HN———C———N
```

Xanthine
2-6-dioxypurine

```
HN———CO
|    |
CO   C—NH
|    ||   \
|    ||    CO
|    ||   /
HN———C—NH
```

Uric acid
2-6-8-trioxypurine

```
NH2
|    CO—NH
|    |    \
OC   |     CO
|    |    /
HN———CH—HN
```

Allantoin

```
HN———CO
|    |
OC   CH
|    ||
HN———CH
```

Uracil
2-6-dioxypyrimidine

```
HN———CO
|    |
OC   C—CH3
|    ||
HN———CH
```

Thymine
5-methyluracil

```
N===C—NH2
|   |
OC  CH
|   ||
HN——CH
```

Cytosine
6-amino-2-oxypyrimidine

```
CH·O
|
HCOH
|
HCOH
|
HCOH
|
CH2OH
```

d-Ribose

```
CH·O
|
CH2
|
HCOH
|
HCOH
|
CH2OH
```

d-Desoxyribose

Uracil takes the place of thymine in ribonucleic acid.

Most nucleic acids have a high molecular weight, therefore they must be made up of several of these units (tetranucleotides), or else they have a more complex structure. Hydrolysis splits up nucleic acid into its constituent nucleotides; phosphatase splits off the phosphoric acid of the nucleotide and leaves a residue (called a nucleoside) formed by the nitrogenous (purine or pyrimidine) base bound to the sugar. The nitrogenous base undergoes deamination while it is still bound to the pentose or after it has been separated. The following are some of the nucleosides:

Adeninepentose, or adenine-nucleoside, or adenosine.
Guaninepentose, or guanine-nucleoside, or guanosine.
Hypoxantinepentose, or hypoxantine-nucleoside, or inosine.

There are three mononucleotides in the tissues that do not form part of nucleic acids but are apparently part of the cell protoplasm. These are

1. Adenosinephosphoric acid, found in muscle. The pentose is *d*-ribose, and it can be united to two or three molecules of phosphoric acid (ADP and ATP). It is important in the chemistry and energetics of muscular contraction.
2. Inosinic acid, also found in muscle. It is formed by hypoxantine, *d*-ribose, and phosphoric acid.
3. Guanylic acid, found in glandular tissue (*e.g.*, the pancreas). It is formed by guanine, *d*-ribose, and phosphoric acid.

These and other similar organic compounds of phosphorus are of great physiologic importance. Thus in cell respiration two dinucleotides act as hydrogen carriers (acceptors and donors). One is diphosphonicotinic amide adenine, or coenzyme I; the other is triphosphonicotinic amide adenine, or coenzyme II.

Digestion and absorption. Pepsin hydrolyzes nucleoproteins into nuclein and protein. Trypsin goes one step further and separates the nucleic acids of nucleoproteins. Enzymes called "nucleases" in the intestinal mucosa and secretion disintegrate nucleic acid into its constituent nucleotides. Other enzymes in the intestinal mucosa, called "nucleotidases" and "nucleosidases," carry hydrolysis still further. Nucleotidases (which are phosphatases) split off phosphoric acid from nucleotides and convert them into nucleosides, which are divided into their respective purine or pyrimidine bases and sugars by nucleosidases. The intestinal mucosa absorbs nucleosides, purine or pyrimidine bases, pentoses, and phosphoric acid. Resynthesis of nucleic acid takes place in the body, but its mechanism is still unknown.

Nucleoprotein catabolism. The disintegration in the organism of the nucleosides absorbed and of nucleoproteins is carried out in a similar way to that just described.

The following diagram gives a summary of the processes of disintegration of nucleic acid:

Nucleic acid
Nuclease ↓

Adenine-nucleotide (adenylic acid) | Guanine-nucleotide (guanilic acid) | Pyrimidine-nucleotides

↓ *Nucleotidase* (phosphatase) ↓

Hypoxanthosine ←* Adenosine | Guanosine →* Xanthosine

↓ *Nucleosidase* (splits off sugar and base) ↓

Adenine | Guanine | Pyrimidine bases

↓ *Adenase* *Guanase* (deamination) ↓

→ Hypoxanthine → Xanthine | Urea + CO_2 + H_2O

Xanthinoxidase (oxidation)

Xanthine
↓ *Xanthinoxidase* (oxidation)
Uric acid
↓ *Uricase*
Allantoin

* Deamination.

Uric acid is a derivative of purine bases. Free purine bases, and even those combined in nucleosides, undergo deamination by specific enzymes; thus adenase converts adenine into hypoxanthine, and guanase converts guanine into xanthine. Hypoxanthine is oxidized into xanthine; the latter is oxidized into uric acid by xanthine-oxidase (xanthine dehydrase).

Autolysis of spleen and liver pulp (both of which contain much nuclear substance) in the presence of blood gives rise to xanthine. If air is

Table 53. Enzymes Acting on Nucleic Acids and Their Derivatives

Enzyme	*Substrate*	*Effect*	*Distribution*
Ribonuclease	Ribonucleic acid	Depolymerization	Pancreas, liver, lungs, leukocytes
Desoxyribonuclease	Desoxyribonucleic acid	Depolymerization	Pancreas, intestinal mucosa, etc.
Nucleotidases (phosphatidases)	Nucleotides	Dephosphorylation	Intestine, liver, etc.
Nucleosidases	Nucleosides	Hydrolysis or phosphorolysis*	Liver, spleen, heart, etc.
Deaminases	Adenine, guanine, adenylic and guanylic acids, adenosine, guanosine, etc.	Hydrolysis at the amine with production of ammonia	Several tissues
Xantinoxidase	Hypoxantine, xantine	Oxidation to uric acid	Liver
Uricase	Uric acid	Oxidation to allantoin	Liver

Source: SUMNER, J. B., and G. F. SOMERS, "Chemistry and Methods of Enzymes," Academic Press, Inc., New York, 1947.

*Phosphorolysis of nucleotides gives ribose-1-phosphate and the nitrogenous base. The process is reversible and therefore also acts as a mechanism of synthesis.

made to bubble through the mass, xanthine is oxidized into uric acid (Hosbaczewski, 1891).

In man and the anthropoids the end product of purine metabolism is uric acid, which is excreted in the urine. In other mammals the greater part of uric acid formed is oxidized by an enzyme, called uricase, into allantoin, which is eliminated in the urine. Oxidation of uric acid can be expressed by the uricolytic index:

$$\frac{\text{Allantoin N} \times 100}{\text{Allantoin N} + \text{Uric acid N}}$$

This index equals 0 in the chimpanzee and 2 in man; in other mammals it varies from 89 to 98. The Dalmatian coach hound has a low index (32); uric acid is the main end product of purine metabolism in this species.

In the dog, uric acid formed in the body or injected is oxidized by uricase in the liver and eliminated as allantoin in the urine. In hepatectomized dogs the uric-acid concentration in blood rises, and urate appears in the urine. If uric acid is then injected, it is no longer excreted as allantoin, but as uric acid.[1] In man there is no uricase in the liver; therefore uric acid formed in the body is eliminated as such. If it is injected, 70 to 100 per cent is eliminated in the urine; the rest is excreted by the intestine or destroyed by bacteria. Uric acid filters through the renal glomeruli and is in part reabsorbed in the renal tubes. The mechanism regulating renal excretion of uric acid is not well known.

The steps in the metabolism of pyrimidine bases are still unknown. The end products of their oxidation are urea, CO_2, and water.

Purine synthesis. Synthesis of purine and nucleoproteins in the organism is proved by many facts: (*a*) purine bases increase in the hen's egg in the process of incubation; (*b*) young mammals are fed almost exclusively on milk, which contains little or no purine, but at that age nuclei are being formed in large quantities; (*c*) a Dalmatian coach hound excreted 100 gm. uric acid in the course of 1 year, during which period it was fed on a diet free from purine (Benedict); (*d*) men who were fed diets without purine for long periods kept in good health and continued to excrete uric acid during the whole period; (*e*) administration of NH_3 labeled with N^{15} is followed by the excretion of purine containing N^{15}; (*f*) glycine is a precursor of purines and of uric acid.

Uricothelic animals. In mammals and amphibians the greater part of nitrogen in protein catabolism is excreted as urea. Uric acid in these species arises exclusively in nucleoprotein and purine catabolism. These animals are known as ureothelic.

In birds and reptiles the principal end product of protein metabolism is uric acid, which arises not only from nucleoprotein, but also from amino-acid catabolism. These animals are called uricothelic. In the goose 60 to 70 per cent of total urinary N is in the form of uric acid, 3 to 5 per cent in urea, and

[1] BOLLMAN, J. L., and F. C. MANN, *Am. J. Physiol.*, **104**, 242, 1933.

10 to 18 per cent in ammonia. Uric acid is formed mainly in the liver. After hepatectomy (Minkowski) uric acid excretion diminishes considerably and there is a marked increase in ammonium lactate. Perfusion of an isolated liver with ammonium lactate gives rise to the formation of uric acid. After the administration of NH_3 labeled with N^{15}, the isotope is found in uric acid.

Excretion of uric acid. Urinary uric acid is said to be endogenous (from nucleoproteins in the tissues) and exogenous (from nucleoproteins in food). Total fasting does not suppress completely the excretion of uric acid, but the amount eliminated increases when there is purine or nucleoprotein in the diet. A normal man on a mixed diet excretes from 0.5 to 1 gm. of uric acid daily, but this figure increases to 1.5 or 2 gm. daily when the diet contains many purine precursors. Purine base is also found in the urine, but only in small quantities (between 15 and 50 mg. daily).

The greater part, perhaps all, of the uric acid eliminated arises from purine or from substances that contain purine. Foodstuffs that contain much nuclear substance (thymus, pancreas, liver, kidney, brain, and meat extract; in a lesser degree, meat, peas, beans, and spinach) increase uric acid excretion. Milk, cheese, cream, and starchy foods contain little or no purine.

A high-protein diet increases the excretion of uric acid in proportion to the caloric value of the diet. Urinary uric acid increases after violent exercise and in fever that is accompanied by a higher rate of protein catabolism. Probably in these cases uric acid arises from adenylic acid in muscle. There is also an increase in uric-acid excretion when there is an excessively active nuclear metabolism, *e.g.*, in leukemia (up to 12 gm. daily), polycythemia, reticulocytosis provoked by liver extracts used in the treatment of pernicious anemia, etc.

Certain beverages contain methylpurines, such as caffeine (in coffee and maté), theophylline (in tea), theobromine (in cocoa, and a little in maté). These are excreted as methylpurines, not as uric acid. Uric-acid excretion is increased by methylpurines only when they act as diuretics, and it is not then a specific effect, since all diuretics increase uric-acid excretion (Thannhauser).

To diminish uric-acid excretion the diet should contain little or no purine and little protein. It should be of low caloric value and should consist mainly of fat and carbohydrate.

Uricemia. There is 2 to 5 mg. per 100 cc. of uric acid in human blood, which is a much higher concentration than that found in the dog. It increases if there is excess production or if it is retained owing to deficient elimination.

The concentration of uric acid in blood is higher in adult men than in women, and gout occurs more frequently in men than in women. When the uric-acid concentration in the blood rises above 5 mg. per cent, the condition is called hyperuricemia. It is observed in (*a*) gout; (*b*) leukemia and polycythemia; (*c*) pneumonia; (*d*) nephritis and the so-called "toxemia of pregnancy." In chronic nephritis blood uric acid rises together with nonprotein nitrogen. Some workers maintain uric acid is one of the first substances to increase in the blood when there is nitrogen retention due to kidney disease.

Gout. In this disease there are (*a*) hyperuricemia, *i.e.*, abnormally high uric-acid concentration in the blood (from 6 to 10 mg. per 100 cc.); (*b*) formation of deposits of sodium urate crystals known as "tophi" in the tissues, particularly in the cartilages of the joints; (*c*) repeated attacks of acute inflammation of the joints. The reason for the deposition of sodium urate crystals is not known. Blood plasma is not saturated with uric acid, and in other diseases (*e.g.*, leukemia, chronic nephritis) there is greater hyperuricemia than in gout without the formation of urate crystals. The uric-acid pool in the organism is increased in subjects with gout, a fact which has been revealed by means of injections of uric acid labeled with an isotope.

Hyperuricemia in gout is not due to excess formation of uric acid, as normal amounts are excreted. Sometimes uric-acid excretion diminishes just before an attack and increases at the beginning of the attack. It would seem that hyperuricemia is due to a selective retention of uric acid by the kidney. Uric-acid clearance at the same level of blood uric acid is below normal in patients with gout. Uric acid injected intravenously into patients with gout is eliminated at a slower rate than in normal subjects. Salicylates, atophan, and cinchophen increase uric-acid elimination in normal and gouty subjects and relieve pain during the attacks of gout.

BIBLIOGRAPHY

Protein Metabolism

Advances in Protein Chemistry
Annual Reviews of Biochemistry

ALBANESE, A. A., "Protein and Amino Acid Requirements of Mammals," St. Luke's Hospital, New York, 1950.

BLOCK, R. J., and D. BOLLING, "The Amino Acid Composition of Protein and Amino Acid Deficiencies," Charles C Thomas, Springfield, Ill., 1948.

DENT, C. E., Blood Amino Acids after Absorption and in Disease, *Biochem. J.*, **44**, 318, 1949; **47**, 223, 1950.

DENT, C. E., *Biochem. Soc. Symposia*, No. 3, Cambridge, New York, 1944.

DENT, C. E., "Liver Disease," Ciba Foundation Symposium, J. & A. Churchill, London, 1951.

GREENBERG, D. M., "Amino Acids and Proteins," Charles C Thomas, Springfield, Ill., 1951.

HAUROWITZ, F., "Chemistry and Biology of Proteins," Academic Press, Inc., New York, 1950.

LEWIS, H. B., Sulfur Metabolism, *Physiol. Rev.*, **4**, 394, 1924.

LEWIS, H. B., Sulfur Metabolism, *J. Nutrit.*, **10**, 99, 1935.

MANN, F. C., The Effects of Complete and of Partial Removal of the Liver, *Medicine*, **6**, 419, 1927.

McCANCE, R. A., The Chemistry of the Degradation of Protein Nitrogen, *Physiol. Rev.*, **10**, 1, 1930.

RITTENBERG, D., Dynamic Aspects of the Metabolism of Amino Acids, *Harvey Lect.*, **44**, 200, 1948–1949.

ROSE, W. C., Essential Amino Acids in Human Nutrition, *Proc. Am. Philos. Soc.*, **91**, 113, 1947.

ROSE, W. C., *et al.*, The Amino Acid Requirements of Man, *J. Biol. Chem.*, **182**, 541, 1950; **188**, 49, 1951; **193**, 605, 1951.

SAYHUN, M., "Proteins and Amino Acids in Nutrition," Reinhold, New York, 1948.

SCHMIDT, C. A., "The Chemistry of Amino Acids and Proteins," Charles C Thomas, Springfield, Ill., 1938 (Suppl., 1944).

SCHOENHEIMER, R., "The Dynamic State of Body Constituents," Harvard University Press, Cambridge, Mass., 1942.

SCHOENHEIMER, R., and D. RITTENBERG, The Study of the Intermediary Metabolism of Animals with the Aid of Isotopes, *Physiol. Rev.*, **20**, 218, 1940.

WASTENEY, S. H., and H. BORSOOK, The Enzymatic Synthesis of Protein, *Physiol. Rev.*, **10**, 110, 1930.

Creatine and Creatinine

BEARD, H. H., "Creatine and Creatinine," Chemical Publishing, Brooklyn, 1943.

HUNTER, A., "Creatine and Creatinine," Longmans, New York, 1928.

ROSE, W. C., Creatine and Creatinine, *Ann. Rev. Biochem.*, **4**, 243, 1935.

TERROINE, E. F., "Créatine et créatinine," Hermann & Cie, Paris, 1938.

Nucleins and Purines

BISHOP, C., *et al.*, Pool Size, Turnover Rate, etc., of Injected Uric Acid, *J. Clin. Investigation*, **30**, 879, 1951.

BRDEREECK, J., Nucleasen, *Ergebn. d. Physiol.*, **7**, 105, 1938.

CHRISTMAN, A. A., Purine and Pyrimidine Metabolism, *Physiol. Rev.*, **32**, 303, 1952.

DANIELLI, J. F., and R. BROWN, "Nucleic Acid," Cambridge, New York, 1947.

DAVIDSON, J. W., "The Biochemistry of Nucleic Acids," Methuen, London, 1950.

GREENSTEIN, J. P., Nucleoproteins, *Advances in Protein Chemistry*, **1**, 210, 1944.

GULLAND, J. M., G. R. BARKER, and D. O. JORDAN, The Chemistry of Nucleic Acids and Nucleoproteins, *Ann. Rev. Biochem.*, **14**, 175, 1945.

HOFFMAN, W. S., Metabolism of Uric Acid and Its Relation to Gout, *J. A. M. A.*, **154**, 213, 1954.

MIRSKY, A. E., Chromosomes and Nucleoproteins, *Advances in Enzymology*, 1944.

SCHLENK, F., Chemistry and Enzymology of Nucleoproteins, *Advances in Enzymology*, **9**, 455, 1949.

STETTEN, DEWITT, On the Metabolic Defect in Gout. *Bull. New York Acad. Med.*, **28**, 664, 1952.

Ammonia

CONWAY, E. J., and R. COOKE, *Biochem. J.*, **33**, 457, 1939,

DELAUNAY, M., and N. POLONOVSKY, *Ann. physiol. physicochim. biol.*, **10**, 695 and 731, 1934.

PARNAS, J. K., and J. HELLER, *Biochem. Ztschr.*, **152**, 1. 1924.

CHAPTER 44

Water Metabolism

The properties of water. Water is quantitatively the main constituent of the body. It is also the universal medium in which all living processes take place; life is not possible without water. The importance of water in biology is due to its remarkable physical and chemical properties. It is the fluid that can dissolve the largest number of substances. Some of these undergo molecular dispersion, others form particles of larger size (micellae of colloidal dispersions), and others are dissociated into ions. Ionic dispersion (*i.e.*, electrolytic dissociation) is made possible by the great dielectric constant of water, which is greater than that of any other liquid. Water plays an essential part in the regulation of body temperature, owing to the following properties:

1. It has a high specific heat, which permits storage of a large amount of heat without a great increase in temperature.
2. It conducts heat, thus assuring its distribution and the establishment of a uniform temperature throughout the body. The circulation is an important factor in the distribution of heat.
3. Its great latent heat of vaporization (540 kg.-cal. is needed to evaporate 1 liter of water) plays a major part in the elimination of heat from the body.

The maximum density of water is at four degrees (centigrade) above its freezing point. This is the reason that ice floats instead of sinking. If the ice on the sea had sunk to the bottom, it would have gradually accumulated and prevented all marine life. Water has been used to establish certain standards, *e.g.*, it is taken as the unit of specific gravity; its freezing and boiling points are taken as 0° and 100° in the centigrade scale of temperature. Water has also a high surface tension. These outstanding properties, unequaled by those of any other fluid, give water not only an essential importance in biology but also great economic and social value.

The amount of water in the body. Concentration of water in the body diminishes throughout life. It is 90 per cent of the body weight in the embryo and 60 to 70 per cent in adults (73 per cent, discounting fat); it is even less in old age. The child not only has a higher concentration of water than the adult; its aqueous equilibrium is also less stable. In the adult, water intake and output are well regulated so as to maintain a constant concentration within narrow limits. When there is excess of water it is rapidly eliminated, and water is retained as soon as its concentration becomes insufficient (Adolph).

Water is found in greatest quantities in the muscles, and next in the skin. Ingested or injected water is taken up mainly by these tissues. In certain abnormal conditions large quantities of water are deposited in the subcutaneous tissue and in the serous cavities (edema).

The amount of water in 100 gm. of tissue is high in the blood, liver, muscles, and skin, and low in cartilage and bone (Table 54-A). It diminishes gradually with age.

Water balance. There are two sources of body water: (*a*) water ingested in food and beverages; (*b*) water formed by the oxidation of hydrogen in the course of metabolism.

The amount of water drunk varies from one individual to another, and from one day to another in the same individual. In temperate climates it varies between 850 and 2,500 ml. per day; on an average 1 ml. per kg.-cal. of heat output. It increases during work and when the temperature of the environment rises. In a hot climate, especially if manual labor or exercise is

performed, water intake can rise up to 2 liters/hr. and 10 or even 13 liters daily; in a hot desert an intake of 6 liters per day is quite usual. Most food contains 60 to 90 per cent water.

An important source of water is the oxidation of foodstuffs or substances stored in the body.

Table 54-A. Distribution of Water in the Tissues

	Weight, kg.	*Percentage of body weight*	*Water*	
			Percentage	*Total* liters*
Body........	70	....	66	46.0
Muscles.....	30	42	76	22.8
Skin........	12	18	72	8.5
Fat.........	13	18	30	3.8
Bones.......	11	16	22	2.4
Blood.......	5.6	8	76	4.2
Liver.......	1.5	2.2	70	1.0

* For a man weighing 70 kg.

Thus for every 100 gm. of fat burned 107 gm. of water is formed; and 55 gm. and 41 gm. for every 100 gm. of starch and protein respectively. This so-called "metabolic water" adds up to 300 or 350 cc. per day (Table 54-B).

Little or no water is absorbed in the stomach. Most of the water ingested is absorbed in the small intestine, and a smaller amount in the large intestine. The latter can absorb up to 80 cc. per hr.; this is utilized in therapeutics for the administration of saline solution by rectum, using the drop method. Isotonic saline solution is also absorbed when injected subcutaneously, intramuscularly, or into a serous cavity. Sodium chloride is rapidly absorbed from the intestine together with water. Other salts, such as magnesium sulfate, are not so easily absorbed; therefore, if they are in a high concentration water will diffuse into the lumen of the intestine and a cathartic effect will result.

Water is eliminated mainly by (*a*) the kidney (1,000 to 1,500 ml. per day); (*b*) the skin (450 to 1,050 ml. daily); (*c*) the lung and air passages (evaporation of 250 to 300 ml. daily); (*d*) the intestine (50 to 200 ml. daily in the feces). In the lactating female a large amount of water is eliminated by milk secretion. Occasionally water is lost in tears, nasal secretion, saliva, secretions of the genital tract, vomiting, or diarrhea.

The amount of water eliminated varies in different circumstances. Thus the volume of urine increases when more salts and nitrogen are eliminated and, generally speaking, when metabolism increases. The output of water is closely related to the intake; *e.g.*, urinary excretion may rise to 10 and even 40 liters per day

Table 54-B. Water Balance of Adult Man, with Moderate Activity at a Temperature of 18 to 20°C.

Intake		*Output*	
Source	*Volume, cc.*	*Route of elimination*	*Volume, cc.*
Beverages......	1,000	Kidney........	1,500
Food..........	1,200	Skin..........	600
Oxidations.....	300	Lung..........	300
		Feces..........	100
Total........	2,500	Total.......	2,500

(polyuria) if large quantities of water are taken (polydipsia). Excessive loss of water causes anhydremia (dehydration); this is seen in cases of digestive fistulas, repeated vomiting, profuse diarrhea, etc.; a typical example is dehydration in patients suffering from cholera.

Cutaneous evaporation of water in basal conditions is 30 to 40 gm. per hr. in a normal subject. This is known as "insensible perspiration" because the water evaporates as soon as it is excreted. When the temperature of the environment rises, sweat secretion increases. The amount of increase depends on the activity of the individual, the clothes worn, and the humidity of the atmosphere. In a hot, dry climate, when performing heavy labor, 5 to 10 kg. per day may be lost by sweating if water is not taken in order to compensate this loss.

Distribution of water. The water in the body is distributed in three separate compartments: (*a*) the blood plasma; (*b*) the tissue fluids—these two constitute the extracellular fluid—(*c*) the intracellular fluid (Fig. 206). There is a continuous exchange of fluid between these compartments; thus in 1 min. a volume of water equal to the water content of plasma enters and leaves the blood vessels.

The amount of water in blood plasma is only about 3 liters, but it is of great importance, because blood plasma is the means of communication between the external environment and the internal environment that surrounds

the cells. It circulates at great speed, and its composition is maintained rigorously constant. Its amount is 45 cc. (37 to 58 cc.) per kg. of body weight; therefore a subject weighing 70 kg. has 3,200 cc. of blood plasma, of which 3,000 is water and 200 protein.

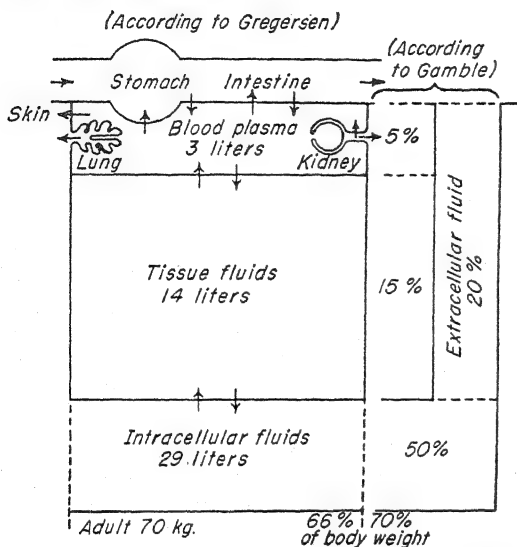



Fig. 206. Diagram of distribution and circulation of water in the organism. (*After Gamble and Gregersen.*)

Extracellular fluid is the sum of blood plasma and tissue fluids. The mineral content of the fluid in these two compartments is similar, but it differs considerably from that of intracellular fluid (Fig. 207). The principal cation in the cells is potassium, and phosphate and protein are the principal anions. The principal cation of extracellular fluid is sodium, and the principal anions are chloride and bicarbonate. The main difference between blood plasma and tissue fluids is in the protein content, because the membrane of the capillaries is permeable to water and salts, but not to protein and other colloids. Proteins in blood plasma, especially serum albumin, exert an osmotic pressure greater than that of proteins in tissue fluids (Starling), but in spite of this the total osmotic pressure is the same in the three compartments, and there are accurate mechanisms that maintain its constancy. Tissue-fluid volume can be determined by several methods. In one method the total sodium or chloride content of the body is measured and then the content in blood plasma is subtracted. In this method it is supposed that sodium and chloride are evenly distributed in all extracellular fluids and that there is none in the cells. There is, however, no proof that NaCl is always in the same concentration in all tissue fluids as in plasma. Moreover, NaCl has been found in small quantities in several tissues.

The method most commonly used for measuring tissue fluids consists in the injection of a known quantity of a substance, such as sodium sulfocyanide (Crandall), which is evenly distributed in all extracellular fluids, but does not penetrate into the cells. The concentration in plasma is then determined, and the extracellular-fluid volume is thus known. Blood-plasma volume is then determined and subtracted from

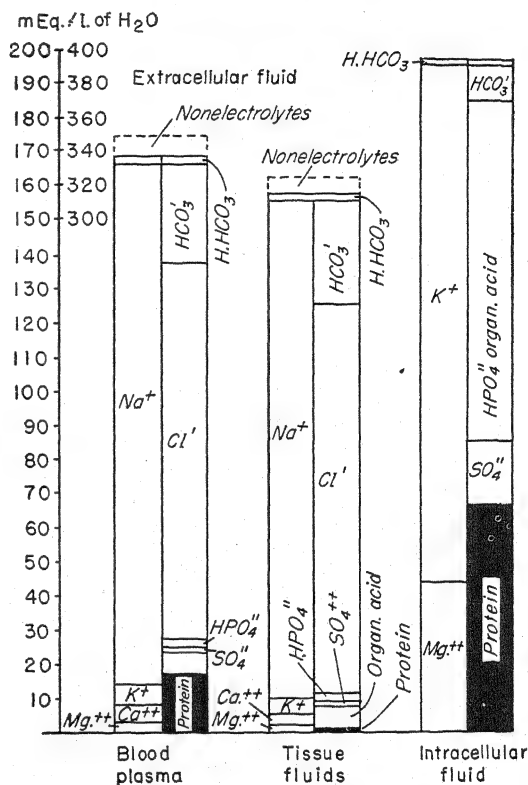



Fig. 207. Principal chemical constituents of the three body-fluid compartments. Columns on the left correspond to total cation and those on the right to total anion concentration. The figures on the scale are milliequivalents per liter of water.

the extracellular volume; the remainder is the tissue-fluid volume. Usually a dye (T 1824, called Evans' blue) is injected together with the sodium sulfocyanide. Other substances (sucrose, inulin, antipyrine, manitol, radioactive Na or Cl, etc.) have also been used for determining

extracellular-fluid volume. None of the methods so far proposed is completely satisfactory and free from error. The normal amount of tissue fluid found with these methods is 180 to 190 ml. per kg. of body weight. The total amount of body weight has been measured by the administration of deuterium in so-called "heavy" water. The distribution of water in the three compartments of the body is given in Table 55.

Table 55. Distribution of Water in the Body

Fluids	*Liters for adult man of 70 kg. (Gregersen)*	*Liters per 100 kg. of body weight*	
		Gregersen	*Gamble*
Extracellular fluids	17	25	20
Blood plasma	3	4.3	5
Tissue fluids	14	18–20	15
Intracellular fluids	29	41	50
Total	46	66	70

The passage of water through the body. Ingested water takes the following path: (*a*) it is absorbed; (*b*) it is transported in blood plasma; (*c*) it diffuses out of the blood capillaries into the tissue fluids; (*d*) it diffuses into the cells; (*e*) it diffuses from the cells back into the tissue fluids; (*f*) it diffuses from the tissue fluids into the blood plasma; (*g*) it is finally eliminated by the kidney, skin, lungs, intestine, etc.

From 3 to 10 liters per day of water is poured into the digestive tract in saliva, in gastric, pancreatic, and enteric juice, and in bile, but most of it is reabsorbed. From 700 to 1,500 cc. of lymph is poured into the blood every 24 hr.

Ingestion of a large quantity of water may cause dilution of the blood, usually of not more than 5 per cent, seldom up to 15 per cent. Dilution occurs in 20 to 40 min. after water has been drunk, and in 50 to 80 min. blood concentration has returned to the initial level. Water is not stored in the blood; it diffuses rapidly into the tissue fluids, and later is eliminated by the kidney.

Not only water but also substances in solution are constantly being interchanged between the three compartments of the body. There is a dynamic osmotic equilibrium between the three compartments which plays an important part in this interchange. The movement of fluid and solutes is caused mainly by

1. The difference in hydrostatic pressure between the capillaries and the tissue fluids. The former is usually greater in the arterial end of the capillary, and water flows out from the blood into tissues.
2. The difference in osmotic pressure between the plasma proteins and the tissue fluids. The former is usually greater and tends to retain water in the blood vessels and to reabsorb it from the tissues, as long as the plasma-protein concentration and the permeability of the capillary membrane remain normal.

Changes in the sodium balance accompany or provoke changes in the water balance. Thus, when sodium is lost, water is also lost and anhydremia occurs. On the other hand, retention of sodium causes retention of water; this is why sodium chloride or bicarbonate solutions are injected into subjects with anhydremia. Changes or disturbance in the potassium balance arise in changes in cell metabolism (Peters).

Potassium and calcium salts are not retained in the body as is the case with sodium salts, because these salts, after having filtered through the glomeruli in the kidneys, are not reabsorbed by the renal tubes; they diminish the reabsorption of water in these tubes, thus having a diuretic effect (see Urinary Secretion). Acidosis also increases diuresis and leads to dehydration. In therapeutics this effect of acidosis is sometimes utilized and ammonium chloride is given in order to provoke acidosis and cause dehydration.

REGULATION OF WATER METABOLISM

Mechanisms. The concentration of water in the body is kept constant within narrow limits. There is an equilibrium between mechanisms that tend to hydrate the organism and others that tend to eliminate excess of water. The body can be hydrated much more rapidly than water can be eliminated.

Water lost by the body is replaced by water absorbed from the digestive tract or, in amphibians and aquatic animals, through the skin. Normally man replaces water by drinking it, and the intake is regulated by the sensation of thirst. Water, however, is also used if it is injected subcutaneously or intravenously. Absorption, transportation, and elimination of water are

regulated by nervous and humoral mechanisms, in which the intestine, liver, and kidney play a leading part; acid-base and ionic equilibriums are also important.

Thirst. The impulse to drink, thus replenishing the water in the body, is induced by the sensation of thirst. It is usually well regulated, and a dog ingests in 2 minutes the exact amount needed to restore water lost (Adolph). Thirst is caused by a decrease in body water due to (*a*) deprivation of fluids; (*b*) loss of water owing to copious sweating, diarrhea, or polyuria; (*c*) hemorrhage, which provokes the passage of water from the tissues to the blood; (*d*) intravenous injection of hypertonic solutions or ingestion of salt, etc.

Thirst has been attributed both to local dehydration of certain tissues and to the general dehydration of the body. General dehydration, by modifying the concentration of salts and water in the blood, is supposed to act on certain nerve centers, thus provoking thirst. Others suppose that dehydration of the cells stimulates afferent nerve endings, which carry impulses to nerve centers and provoke thirst.

Cannon[1] and Gregersen[2] attribute thirst to drying of the mucosae of the mouth and pharynx, especially of the fauces, owing to a decrease in the secretion of saliva. Decrease in salivary secretion has been observed in anhydremia, hemorrhage, injection of hypertonic saline solution, and other conditions in which there is dehydration. In these cases thirst disappears if the mucosae of the mouth and pharynx are wetted, but thirst soon reappears because this water rapidly evaporates and is not replaced, as salivary secretion does not increase while the blood remains dehydrated.

The theory that attributes thirst to general dehydration of the tissues is supported by Claude Bernard's experiments on horses with a fistula in the esophagus and dogs with a gastric fistula. These animals drank continuously, stopping only to rest for a few moments; their thirst was never satisfied, because the water drunk was lost through the fistula. If 500 to 1,000 cc. of water was introduced by a catheter into the stomach, thus preventing its loss, thirst was soon appeased. Thirst can also be slaked by intravenous injection of water (Orfila) or isotonic saline. An increase in the osmotic pressure of the blood plasma, such as is observed in diabetic hyperglycemia or after hypertonic salt injection, provokes thirst. Cell dehydration would be the cause of the desire for water, according to Gilman, Dill, and others. So far this theory has not been demonstrated by direct experiments, and it does not explain all the facts observed.

[1] Cannon, W. B., *Proc. Roy. Soc., London, s.B.*, **90**, 283, 1918.

[2] Gregersen, M., in MacLeod, "Physiology in Modern Medicine," 9th ed., Mosby, St. Louis, 1941.

Bellows's[1] observations on dogs with a fistula in the esophagus tend to conciliate the theories that attribute thirst to a local cause with those which consider that it is caused by a general dehydration of the tissues. By wetting the mouth and pharynx, thirst was appeased immediately for a short time. By introducing water into the stomach, thirst was appeased after a certain delay, and the effect lasted for some time.

The importance of certain nerve centers will be discussed when considering the neurohypophysis and the hypothalamus.

Changes in water and electrolyte concentration in the three compartments (blood plasma, tissue fluids, and intracellular fluids), occurring in thirst due to different causes, are not yet sufficiently well known for their significance with respect to the sensation of thirst to be understood.

The part played by different organs in the regulation of water metabolism. Water is absorbed mainly in the small intestine and to a lesser degree in the large intestine. The liver can store water, and certain disturbances in the circulation of the liver cause a decrease in diuresis. In some forms of hepatosis and cirrhosis there is a delay in diuresis after the ingestion of water (opsiuria). An antidiuretic factor has been found in the plasma and urine in patients with a diseased liver; the effect of this organ on water metabolism is, therefore, not exerted only by physical means.

The kidney plays an important part in water metabolism. Approximately 170 liters of fluid passes daily through the glomeruli, and 168.5 liters is reabsorbed in the renal tubes; 1.5 liters is eliminated as urine. Excess water is eliminated by the kidney. Moreover, the kidney eliminates selectively water or salts so as to keep constant the osmotic pressure in the body fluids.

The neurohypophysis secretes an antidiuretic hormone which regulates the absorption of

[1] Bellows, R. T., *Am. J. Physiol.*, **125**, 87, 1939.

water in the distal renal tubes. Secretion of this hormone is under the continuous control of the supraoptic nuclei, which send fibers in the tractus supraopticohypophysis through the hypophyseal stalk to the neurohypophysis. If the neurohypophysis is removed, or if it is denervated by section of the tractus supraoptico-hypophysis or by destroying the supraoptic nuclei, the antidiuretic hormone is no longer secreted and water is reabsorbed by the distal renal tubes at a lower rate than in normal animals. There is marked polyuria; 5 to 10 liters, and in some cases up to 40 liters, of urine is passed daily. The condition is known as diabetes insipidus. Such considerable loss of water causes an intense thirst and the ingestion of large quantities of water (polydipsia). If the subject abstains from drinking water, his blood concentration increases (anhydremia). Administration of posterior hypophyseal extract increases reabsorption by the renal tubes, diminishes diuresis, and causes the signs of diabetes insipidus to disappear. In normal subjects posterior lobe extracts diminish spontaneous diuresis and diuresis following the ingestion of water.

Corticoadrenal hormones regulate the reabsorption of sodium by the renal tubes. Adrenalectomy causes loss of sodium in the urine, accompanied by chloride and water (polyuria). Potassium, on the contrary, is reabsorbed at a greater than normal rate. These abnormalities in renal excretion cause a decrease in plasma sodium and chloride and in extracellular-fluid volume, with a decrease in blood-plasma volume (hemoconcentration). At this stage of adrenal insufficiency there is no longer polyuria but oliguria. Ingested water is eliminated very slowly in adrenal insufficiency, a fact which has been used to establish the diagnosis of this condition. Administration of certain corticoadrenal hormones increases reabsorption of sodium, chloride, and water and decreases reabsorption of potassium in the renal tubes. Therefore in cases of adrenal insufficiency treated with these hormones excess excretion of sodium and chloride in the urine ceases and Na and Cl concentration in plasma increases and that of K decreases. Desoxycorticosterone has these effects and maintains life in adrenalectomized animals. An excess of this hormone causes polyuria and polydipsia. If the diet contains large quantities of NaCl at the same time as this hormone is given, salt and water are retained in the extracellular fluids, and plethora, edema, hypertension, and cardiac insufficiency may follow.

Estrogenic hormones can provoke some retention of NaCl and water. This is observed in the sexual skin of female monkeys. In women there is frequently a slight periodic premenstrual increase in weight, and sometimes edema. If the ingestion of salt is restricted this increase is not observed.

The thyroid plays an important part in water metabolism. In thyroid insufficiency water is retained together with protein and infiltrates the skin (myxedema). Administration of thyroid causes an increase in diuresis and loss of body weight during a few days, particularly marked in myxedematous subjects, until an equilibrium is established. Polyuria is also observed when the thyroid is stimulated by the thyrotrophic hormone of the hypophysis.

DISTURBANCES IN WATER METABOLISM

Water metabolism is disturbed in the following circumstances:

1. When water is lost in relatively larger amounts than salt, as occurs in deprivation of water, copious sweating, and diabetes insipidus.
2. When the loss of salt is relatively greater than the loss of water, as in acute adrenal insufficiency, diabetic acidosis, and copious loss of fluid through the stomach or intestine.
3. When salt and water are lost in equivalent amounts, as in fasting.
4. When salt is retained in excess of water, as when sea water is drunk.
5. When water is retained in excess of salt, as in water intoxication (see page 476).
6. When salt and water are retained in equivalent amounts, as in edema.

Edema is due to excessive accumulation of tissue fluid, sometimes accompanied by effusion in the serous cavities. It is observed in cardiac decompensation, in cases of obstruction of the venous circulation, in cirrhosis of the liver, and in malnutrition (hypoproteinemia, some forms of kidney disease, etc.).

Dehydration. In true dehydration the water content of the body is diminished, *i.e.*, there is a deficiency of water. It should be distinguished

from salt deficiency, although the two conditions are frequently associated.

Subjects deprived of food but not of water survive much longer than those deprived of food and water. Luciani mentions a political prisoner who went on total food strike and drank no water; he survived 17 days. After a few days hunger was no longer felt, but he suffered from an intense thirst until shortly before he died. In the hot, dry climate of a desert, death occurs after 36 to 72 hr. of total water deprivation. Experimental data have established that a loss of water equivalent to 10 per cent of the body weight causes serious disturbances; a loss of 20 to 25 per cent is usually fatal.

The following signs are observed in dehydration: loss of weight; thirst; sunken eyes with softening of the eyeballs; dry lips, tongue, and fauces. The skin is dry; it loses its elasticity and becomes wrinkled. Nonprotein nitrogen Na and Cl concentration in plasma increases; K concentration decreases. There is frequently acidosis. Urinary volume is considerably diminished; the urine is of high density and has a high sodium chloride content. In the final stages there is circulatory collapse.

Acidosis provokes loss of water and anhydremia, which, if sufficiently severe, causes circulatory failure. This occurs in diabetic coma; therefore in the treatment of this condition dehydration must be taken into account. Reciprocally, dehydration (such as occurs in copious diarrhea) can lead to acidosis or to its increase if it is already present.

Sodium chloride deficiency also causes dehydration. There is loss of weight, a dry tongue, sunken eyes with low intraocular pressure, an increase in plasma nonprotein nitrogen, etc. There are, nevertheless, certain symptoms that distinguish this condition from simple dehydration. There is loss of appetite, intense headache, lethargy and even mental confusion, dizziness, and sometimes loss of consciousness; frequently there are painful muscular cramps. All the symptoms of dehydration are present in a severe form, and circulatory collapse is easily provoked. The urine contains little sodium chloride and is of low density, contrasting with the urine of simple dehydration, which has a high density and NaCl concentration.

Salt deficiency is observed when intense manual labor is performed in a hot climate, because NaCl is lost in sweat. Water is usually drunk in sufficient quantities to replace the loss of water, but sodium chloride is not replaced. After a few days, weight is lost, there is anhydremia, and blood NaCl concentration is decreased. The signs of deficiency disappear as soon as NaCl is ingested; they can be prevented by adding 0.1 per cent NaCl to the drinking water. Water alone does not improve the subjects; on the contrary, copious water drinking increases sweating and the loss of NaCl, thereby increasing the seriousness of the condition.

Simple anhydremia is rapidly cured by drinking water. In other cases of anhydremia, sodium chloride (0.7 to 0.9 per cent) should also be given, and when there is hypochloremia, KCl and sodium bicarbonate should be added. Mortality in infantile diarrheas is considerably diminished by adequate treatment with salts and water (see Chap. 45, Mineral Metabolism).

Water intoxication. If the kidney is functioning normally, excessive ingestion of water is well tolerated, because the excess is rapidly eliminated in the urine, and hydremia is prevented or occurs in only a mild form. If the kidney is diseased, copious water drinking causes a disturbance known as water intoxication. There are headache and general malaise, vomiting, painful muscular cramps, an increase in body weight, and a rise in blood pressure. In the dog, water intoxication has been provoked by giving large quantities of water (50 cc. per kg.) by stomach tube and injecting antidiuretic posterior hypophyseal extract. In these animals there are asthenia, restlessness, salivation, vomiting, diarrhea, tremor, clonic muscular contractions and convulsions, lethargy, coma, and finally death. At autopsy cerebral edema is frequently found.

Excessive ingestion of water, especially if antidiuretic hormone is given, causes convulsive attacks in epileptic patients.

In pregnancy retention of water may lead to eclampsia with fatal convulsions.

BIBLIOGRAPHY

Adolph, E. F., The Metabolism and Distribution of Water in Body and Tissues, *Physiol. Rev.*, **13**, 336, 1933.

Adolph, E. F., "Physiological Regulations," Cattell and Company, Incorporated, Lancaster, Pa., 1943.

Adolph, E. F., "Physiology of Man in the Desert," Interscience Publishers, Inc., New York, 1947.

Dill, D. B., "Life, Heat and Altitude," Harvard University Press, Cambridge, Mass., 1938.

Edelman, I. S., J. M. Olney, A. H. James, L. Brooks, and F. Moore, Body Composition. Studies in the

Human Being by the Dilution Principle, *Science*, **115**, 447, 1952.

GAMBLE, J. L., "Chemical Anatomy, Physiology and Pathology of Extracellular Fluid," Harvard Medical School, Boston, 1942.

GAUNT, R., J. H. BIRNIE, and W. J. EVERSOLE, Adrenal Cortex and Water Metabolism, *Physiol. Rev.*, **29**, 281, 1949.

GREGERSEN, M. J., in MacLeod, "Physiology in Modern Medicine," 9th ed., Mosby, St. Louis, 1941.

HEVESY, G., and C. F. JACOBSEN, Rate of Passage of Water through Capillary and Cell Walls, *Acta physiol. Scandinav.*, **1**, 11, 1940.

JANOWITZ, H., and M. I. GROSSMAN, "Hunger and Appetite," precirculated mimeographed copy.

MARRIOTT, H. L., "Water and Salt Depletion," Charles C Thomas, Springfield, Ill., 1949.

MARX, H., "Der Wasseraushalt des gesunden und kranken Menschen," Springer, Berlin, 1935.

MCCANCE, R. A., Salt Deficiency, *Lancet*, **1**, 643, 704, 765, and 823, 1936.

PETERS, J. P., "Body Water," Charles C Thomas, Springfield, Ill., 1935.

PETERS, J. P., Water Exchange, *Physiol. Rev.*, **24**, 491, 1944.

PETERS, J. P., Water Balance in Health and Disease, in Duncan, "Diseases of Metabolism," 2d ed., Saunders, Philadelphia, 1947.

PETERS, J. P., and D. D. VAN SLYKE, "Quantitative Clinical Chemistry," 2d ed., vol. 1, Interpretations, Williams & Wilkins, Baltimore, 1946.

PINSON, E. A., Water Exchanges and Barriers as Studied by the Use of Hydrogen Isotopes, *Physiol. Rev.*, **32**, 123, 1952.

SCHMIDT-NIELSEN, K., and B. SCHMIDT-NIELSEN, Water Metabolism and Desert Animals, *Physiol. Rev.*, **32**, 135. 1952.

CHAPTER 45

Mineral Metabolism

WATER AND SALTS are the inorganic or mineral constituents of the body. They fulfill specific functions in the cells and are indispensable for life. Each type of cell has a particular mineral composition and water content, which is kept constant within narrow limits. Some mineral substances, *e.g.*, sodium chloride, are found in relatively large quantities; they take part in maintaining the osmotic pressure and acid-base equilibrium. Others are found in very small quantities but may be of great physiologic importance; *e.g.*, iodine, which is indispensable for normal thyroid function.

Biologic activity of minute quantities of metals (*oligodynamic effect*) was first demonstrated by Raulin in his classic experiments on the growth of *Aspergillus*.

Oxygen, carbon, hydrogen, and nitrogen in water and organic substances form approximately 95 per cent of the human body. Calcination reduces the body to ashes, which weigh 4.5 to 5 per cent of the total body weight. According to Bertrand,[1] 36 chemical elements have so far been found in animal organisms; 29 are always present, 7 are occasionally found. Among those always present are 13 nonmetals (C, H, O, N, S, P, Cl, F, Br, I, B, As, Si), and 16 metals (Na, K, Ca, Mg, Fe, Zn, Cu, Ni, Co, Mn, Al, Pb, Sn, Mo, Va, Ti). Those occasionally found are Rb, Cs, Li, Ba, Sr, Ag, and Cr (Table 56).

Table 56. Principal Constituent Elements of the Human Body

Constituent element	Percentage of body weight	Constituent element	Percentage of body weight	Constituent element	Percentage of body weight
Oxygen	65	Calcium	2	Magnesium	0.05
Carbon	18	Phosphorus	1	Iron	0.005
Hydrogen	10	Potassium	0.35	Manganese	0.0003
Nitrogen	3	Sulfur	0.25	Copper	0.00022
O, C, H, and N	96	Sodium	0.24	Iodine	0.00004

Data taken from Volkman, Sherman, Widdowson, and McCance, *et al.*

Mineral substances are found in extracellular and intracellular fluids; their concentrations are maintained constant by means of regulatory mechanisms in which the kidneys, lungs, erythrocytes, and tissues play a part. They are important factors in the following functions:

1. Osmotic equilibrium.
2. Acid-base equilibrium.
3. Ionic equilibrium (Na^+, K^+, Ca^{++}, Mg^{++}), which is an important factor in the excitability of nerve and muscle, capillary and cell permeability, cardiac functions, etc.
4. Water metabolism and exchange of water between blood plasma and extracellular and intracellular fluids.
5. Enzymatic systems (Mg^{++} of phosphatase, carboxylase, cozymase; Mn^{++} of arginase, cholesterinase, Zn^{++} of carbonic anhydrase).
6. Respiratory function, by means of the part played in carrying CO_2 from the cells to the lung, and of Fe in ferroporphyrins (hemoglobin, myoglobin, cytochrome).

[1] BERTRAND, G., Congrès Chim. Biol., Paris, 1937.

7. As constituents of tissues, especially bone (mainly calcium and phosphate), to which they give consistency in life and duration after death.
8. Thyroid function. Iodine forms part of the thyroid hormone which regulates oxygen consumption and heat production.

Excretion. Sodium, chloride, and potassium are eliminated mainly by the kidney, which excretes 95 per cent of the potassium. These elements are ultrafiltrated into the glomerular fluid and are then in great part reabsorbed by the renal tubes. Approximately 180 liters of fluid, containing 1,200 gm. of NaCl, is filtered daily through the glomeruli, but most of this is reabsorbed and only 1.5 liters of water with 5 to 10 gm. of NaCl is eliminated in the urine. The kidney is the main organ of regulation of electrolyte concentration and water in the body; therefore it also regulates osmotic pressure, and ionic and acid-base equilibriums. Acids combined with sodium and other bases in blood plasma are combined with ammonium in the kidney and excreted in the urine as ammonium salts, while most of the base is reabsorbed and returned to the plasma.

Table 57. Electrolyte Concentration in Body Fluids of Man in Milliequivalents* per Liter

Body fluid	Na^+	K^+	Ca^{++}	Mg^{++}	Cl^-	HCO_3^-	$PO_4^{=}$	SO_4^-	*Organic acid*
Intracellular fluid	40	118	3	20	21	10	110	1	...
	37	112	..	..	25	...	...	...	...
Tissue fluid	138	5	5	3	108	27	2	1	6
Blood plasma	142	5	5	3	103	27	2	1	6

* The advantage of expressing electrolyte concentration in milliequivalents is explained on page 9.

DISTRIBUTION OF ELECTROLYTES IN THE BODY FLUIDS

The electrolyte content of body fluids in the three compartments is given in Table 57.

Tissue fluid varies considerably in the different tissues, but the composition of blood plasma is maintained constant. Blood plasma has a relatively high concentration of Na and Cl; this is also true of tissue fluids, which, however, have a lower protein concentration. Intracellular fluid has a high concentration of K^+, Mg^{++}, $PO_4^{=}$, and protein, and low concentrations of Na^+, Cl^+, and HCO_3^-.

METABOLISM OF SODIUM, CHLORIDE, AND POTASSIUM

Requirement and ingestion. Complete deprivation of any of these three elements causes growth to cease and eventually the animals die. Deficiency in human beings is seldom seen because the diet usually contains them in adequate amounts. There is a daily ingestion of 5 to 15 gm. of sodium chloride, but a normal balance can be maintained with 2 to 3 gm. daily. A normal diet provides 4 to 6 gm. of sodium and 2 to 4 gm. of potassium daily. Special diets poor in either of these elements can be prepared easily, as there are tables which give the sodium and potassium content of different foodstuffs. During growth there is greater need of potassium because large amounts are retained in the protoplasm, especially in young cells.

Elimination of sodium chloride from the diet causes a decrease in sodium chloride concentration in plasma. The normal average level is 600 mg./100 cc. (580 to 620 mg. per cent); when this falls to 560 mg. per cent, urinary excretion of sodium chloride diminishes and is soon reduced to traces, or it may even cease completely. Simultaneously there is a decrease in sodium chloride and hydrochloric acid excreted in the gastric juice. If, on the contrary, an excess of sodium chloride is ingested, it is eliminated in one to two days. Sodium chloride is retained only when there has been a deficiency of this salt or when there is anhydremia or a tendency to form edema. Potassium is eliminated at a faster rate than sodium. The use of isotopes has shown that the passage of sodium through the organism takes place at a rapid rate. A normal subject exchanges about half the total sodium content of the body in approximately 15 days.

Sodium is the cation in highest concentration

in plasma; it is in equilibrium with Cl^-, HCO_3^-, fixed acids and protein. Sodium combined with bicarbonate is available for neutralization of acids and is therefore known as the alkaline reserve of the plasma.

The concentration of all bases in plasma is 155 mEq./liter. Sodium constitutes 92 per cent of total base, and Cl^- makes up two-thirds of the anions.

Changes in acid-base equilibrium are caused by a deficiency in one of the ions in extracellular fluids. Thus Cl^- deficiency causes alkalosis and Na^+ deficiency causes acidosis. For example, loss of Cl^- by repeated vomiting produces alkalosis; loss of alkali by a pancreatic, biliary, or intestinal fistula causes acidosis.

Sodium chloride is the main factor in maintaining osmotic equilibrium. Retention of Na and Cl causes the retention of water, and retention of water causes Na and Cl to be retained. Excretion of Na and Cl is accompanied by the excretion of water, and the elimination of large quantities of water is accompanied by excretion of Na and Cl. Thus, when a solution of glucose is injected, glucose is metabolized and the excess water eliminated carries sodium chloride with it. A decrease in extracellular sodium causes water to pass into the cells which become swollen.

A normal Na and Cl concentration in the blood plasma is of great importance to the organism. Adrenalectomized animals survive as long as NaCl is given them in sufficient amounts to compensate the loss of Na; as soon as the salt treatment is discontinued they fall into acute adrenal insufficiency and die. In Addison's disease (adrenal insufficiency in man) salt treatment also prevents and cures acute adrenal crises and prolongs life.

Subjects performing heavy labor in a hot environment sweat profusely and lose NaCl. They suffer from tiredness, insomnia, and painful muscular cramps. These symptoms are prevented by adding 0.1 to 0.2 per cent NaCl to the drinking water or by the ingestion of salt tablets.

Edema consists in the accumulation of sodium chloride and water in the subcutaneous tissue; therefore Na, Cl, and water are retained in the body. In these cases ingestion of NaCl causes an increase in edema owing to further retention of salt and water. Elimination of NaCl, on the contrary, causes loss of water and diminishes edema. Most diuretic substances decrease reabsorption of NaCl and water in the renal tubes, thus increasing urinary elimination of these substances.

Large amounts of sodium, chloride, and potassium can be lost and anhydremia provoked by (*a*) profuse sweating, polyuria, diarrhea, vomiting, loss of fluid from fistulas in the digestive tract; (*b*) acidosis due to excess renal excretion of salts and water; (*c*) rapid transudation of water and salts in surgical and traumatic shock, pulmonary edema, acute intoxication with mercury, etc. (see Chap. 44, Water Metabolism).

Ingestion of ion-exchange resins which adsorb sodium and potassium may cause a considerable loss of these elements through the digestive tract. This principle has been used in therapeutics for reducing edema, etc., but care should be taken that not too much potassium is lost.

Corticoadrenal hormones control reabsorption of sodium in the renal tubes. In acute adrenal insufficiency this reabsorption diminishes, therefore renal excretion of sodium, chloride, and water increases and that of potassium diminishes. Sodium and chloride concentration in plasma diminishes, and potassium concentration increases. Corticoadrenal extract and desoxycorticosterone reverse these conditions: Na and water are retained, and renal excretion of K increases; in blood plasma, Na and Cl concentrations return to normal and that of K decreases. An excess of desoxycorticosterone in patients with adrenal insufficiency and even in normal subjects may cause excess retention of Na and Cl, which provokes retention of water and subsequent edema, hypertension, and cardiac dilatation (see Chap. 54, The Adrenal Glands).

In heart insufficiency there is congestion of the kidneys, which, according to some authors, increases glomerular filtration and tubular reabsorption of Na. Water would also be reabsorbed, thus increasing plasma volume and venous and capillary pressure. Finally, transudation of fluid from plasma to the tissues would take place, with formation of edema (Stead). According to others, mechanical factors (*i.e.*, increased venous and capillary pressure) are the main cause of edema in cardiac disease.

Salt-poor diets, *e.g.*, Kempner's rice diet, have been used in the treatment of hypertension.

Ingestion of potassium salts causes diuresis because elimination of K in the urine diminishes reabsorption of Na, Cl, and water by the renal tubes.

Mercurial diuretics diminish tubular reabsorption of Na and Cl which carry water away with them, thus causing an increase in urinary volume.

In certain forms of acute nephritis glomerular filtration diminishes, Na is retained and there is a tendency to form edema.

The role of potassium. Potassium is the predominant cation in the cells. It is the main factor in osmotic equilibrium and acts as available base in acid-base equilibrium; it is a constituent of essential enzymatic systems and plays a part in electrical phenomena. Changes in potassium concentration have a marked influence on most of the functions of the cells.

A 70-kg. man has approximately 170 gm. of potassium, of which 9 gm. is in the blood (only 0.6 gm. in plasma), and 3 gm. in tissue fluids. Potassium is found in relatively large amounts in cells, *e.g.*, in muscle and liver cells and in erythrocytes (400 mg./100 ml.), and in much smaller quantities in extracellular fluids, *e.g.*, 17 mg./100 ml. in blood plasma if the erythrocytes are separated without delay. This high intracellular concentration of potassium is maintained against osmotic and electrical forces by mechanisms which are not yet well understood. Studies with isotopes of potassium have shown that it is continuously entering the cells and leaving them and that a dynamic equilibrium maintains a constant concentration.

The concentration of potassium in blood plasma is small but of great physiologic importance, because it is in ionic equilibrium with Na, Ca, and Mg, so that when it increases or diminishes it has a marked effect on many functions of the cells. Variations of plasma-potassium concentration are therefore of great interest, and they can be rapidly and accurately determined by means of chemical analysis or the flame photometer. The level of potassium in plasma, however, cannot be taken as reflecting its concentration in the cells, because these may be depleted of potassium although the plasma level remains within normal range.

When potassium leaves the cells it is replaced by sodium. There is a transitory outward migration during activity in muscle and nerve, and a rapid return during the phase of recovery. Glycogenolysis is also accompanied by a loss of potassium, and glycogenesis by accumulation within the cells. Hypoxia or anoxia and other agents which damage the cell provoke loss of potassium. Diets with low potassium content cause a marked and sustained decrease in potassium concentration of muscle. Potassium is accumulated during normal growth and during growth provoked by testosterone and somatotrophin.

Potassium deficiency may be due to

1. Insufficient amount in the diet.
2. Loss in digestive secretions owing to vomiting, diarrhea, and fistulas. Up to one-quarter the total amount in the body may be lost in cases of profuse diarrhea.
3. Increased renal secretion as in diabetic acidosis, abundant diuresis caused by injection of glucose or saline solutions, or excess administration of corticoadrenal hormones (desoxycorticosterone).
4. Increased protein catabolism.
5. Injecting fluids without potassium (glucose, sodium chloride solutions, etc.), which can provoke or increase deficiency in cell potassium.

Potassium deficiency may cause a fall in plasma potassium; this also occurs when the cells take up potassium at a rapid rate, *e.g.*, after the injection of glucose, insulin, or small doses of adrenaline and in familial periodic paralysis. An excess of corticoadrenal hormones provokes loss of K and Cl, consequently alkalosis with hypopotassemia and hypochloremia; this is seen in some cases of hypercorticalism, Cushing's disease, and after injection of adrenocorticotrophin. A slight decrease in plasma potassium has been observed in the course of hypophyseal insufficiency in the dog, and a transitory fall during anesthesia.

A fall in the concentration of plasma or intracellular potassium if sufficiently marked will cause loss of appetite, meteorism, lethargy, muscular weakness leading to paralysis, dilatation of the heart, hypotension, and pulmonary edema. A prolonged fall may produce foci of necrosis in the myocardium and other tissues. The electrocardiogram shows a low and wide T wave, lengthening of the QT and sometimes the PR intervals, and ST depression when the plasma concentration falls to 10 to 12 mg./100 ml. (2.5 to 3 mEq./liter).

Potassium deficiency is treated by the administration of potassium, preferably by mouth. If it must be injected, it should be by slow intra-

venous infusion of balanced solutions with more Na than Cl, taking care to remain well below toxic levels. Treatment by infusion of balanced solutions has lowered considerably the mortality in cases of diabetic coma and infantile diarrheas.[1]

An increase in plasma potassium to 30 mg./100 ml. or more has toxic effects which may cause death. Plasma potassium increases in the following cases: (*a*) when an excess of K is incorporated into the organism; (*b*) when there is renal insufficiency (deficient elimination of K) and the intake is not reduced or suppressed; (*c*) when there is an increase in protein catabolism, cytolysis or cell injury as in hemolysis, trauma, hypoxia, ischemia, or shock; (*d*) when there is strong, sustained, and widespread muscular contraction, *e.g.*, tetanus, convulsions (during contraction K goes out of the muscle fibers, and returns during recovery); (*e*) when the sympathetic system is stimulated or adrenaline is injected or secreted in sufficiently large quantities (K is released from the liver to return during the period of recovery—sympathico-adreno-hepatic mechanism); (*f*) in dehydration owing to an increase in concentration of extracellular fluid; (*g*) in severe uncompensated corticoadrenal insufficiency, but not in compensated insufficiency (renal excretion of K diminishes and K increases in extracellular fluids but diminishes within the cells).

When plasma potassium increases, the electrocardiogram shows a high and narrow T wave, lengthening of PR interval, and depression of ST; P may disappear, and the QRST complex may show abnormalities. The heart stops in diastole when K concentration in plasma rises to 39 mg./100 ml. (10 mEq./liter).

Potassium can be removed from the organism by hemodialysis or by oral administration of ion-exchange resins.

CALCIUM METABOLISM

Functions of calcium. Calcium makes up about 2 per cent of the body weight; most of it (98 per cent) is in the bones. It has many functions, among which the following are the most important: (*a*) it is an essential factor in the development of bones and teeth, which owe their consistency to calcium salts; (*b*) it is an important factor in the regulation of membrane permeability;[1] (*c*) it is essential for nearly all the functions of the cell; (*d*) in ionic equilibrium with Na, K, and Mg, it regulates neuromuscular excitability; (*e*) it is a necessary factor in blood clotting and milk coagulation; (*f*) it plays a part, although only a secondary one, in the regulation of water metabolism and acid-base equilibrium. Calcium is closely associated with phosphorus in the process of ossification. Variations in blood-calcium concentration are generally in the opposite direction to those in blood-phosphorus concentration. Thus hypocalcemia is usually accompanied by hyperphosphatemia, and hypercalcemia by hypophosphatemia.

The important factor in the regulation of cell functions and neuromuscular excitability is the concentration of calcium ion, not of total calcium. Calcium ion depresses excitability, but its effect depends on the concentration of other ions. Monovalent ions (Na^+ and K^+) have the opposite effect of bivalent ions (Ca^{++} and Mg^{++}); hydrogen ion has the same effect as bivalent ions. Excitability depends on an equilibrium between these ions, which can be expressed as follows:

$$K = \frac{(Na^+) + (K^+)}{(Ca^{++}) + (Mg^{++}) + (H^+)}$$

Excitability is depressed when the value of the constant K diminishes and it becomes higher as K increases.

Hypocalcemia (*i.e.*, total blood Ca concentration below normal) is accompanied by a decrease in Ca ion concentration; the ionic equilibrium is disturbed, and excitability increases. Neuromuscular hyperexcitability causes an increase in muscle tonus, muscular spasms, tremor, and in some cases convulsions. This condition is known as tetany. Hypercalcemia, on the contrary, depresses excitability in the peripheral tissues and in the nerve centers.

Tetany. Neuromuscular hyperexcitability is produced in several ways: (*a*) fall in blood calcium; (*b*) injection of phosphate, which provokes a fall in blood calcium; (*c*) injection of citrate, which causes a decrease in Ca ion; (*d*) alkalosis due to excessive ingestion of bicarbonate or loss of Cl in repeated vomiting (gastric tetany); (*e*) alkalosis provoked by hyperpnea; (*f*) decrease in magnesium; (*g*) certain nerve poisons, such as

[1] Darrow, D. C., *Bull. New York Acad. Med.*, **24**, 147, 1948.

[1] Calcium diminishes membrane permeability; it also diminishes hydration of colloids, and renders gels firmer; potassium and sodium have the opposite effects.

guanidine and ergot. Table 58 summarizes the condition of the blood in different types of tetany.

Ingestion and requirement. Calcium forms part of a normal diet. Milk (0.15 per cent), cheese (0.5 to 1 per cent), and green vegetables, if they have not been deprived of Ca by boiling, are the foodstuffs that contain the most calcium. Certain mineral waters also have a high calcium content. There is a very large store of calcium and phosphorus in the bones. Ca and P are continuously entering and leaving this store, and there are precise mechanisms that regulate this interchange and maintain an equilibrium in calcium metabolism. This equilibrium is dependent on (*a*) calcium ingested; (*b*) calcium excretion; (*c*) regulatory factors, the most outstanding of which are vitamin D and the parathyroid hormone. The concentration of calcium in the blood is not a reliable index of calcium metabolism, which can be estimated only by measuring the calcium balance (*i.e.*, calcium intake and output) and examining the condition of the bones by means of radiographs, histologic examination, and chemical analysis.

The minimum daily calcium requirement is 0.45 gm., according to Sherman; but the normal diet for an adult should contain approximately twice this amount, *i.e.*, 1 gm. Ca requirement increases in pregnancy and lactation, owing to the formation of the bones in the fetus and the secretion of milk; pregnant women should receive 1.5 gm. daily and lactating women 2 gm. daily. If there is an insufficient Ca intake the mother's bones are decalcified, and even so the child's bones and teeth may be deficiently calcified and suffer permanent damage. Children should be given 1 gm. of Ca daily between the ages of 1 and 12 years. Girls should have 1.3 gm. from 13 to 15 years and 1 gm. from 16 to 20 years; boys, 1.4 gm. from 13 to 20 years.[1]

Table 58. Types of Tetany

Cause	*Concentration in blood plasma*				*pH*
	Ca	*Ca^{++}*	*PO_4*	*CO_3H^-*	
1. Calcium deficiency					
a. Rickets and osteomalacia*	Reduced	Reduced	Reduced	Normal	Normal
b. Insufficient Ca intake	Reduced	Reduced	Increased	Normal	Normal
c. Oxalate	Reduced	Reduced	?	Normal	Normal
d. Intestinal (sprue, celiac disease)	Reduced	Reduced	?	Normal	Normal
e. Parathyroid	Reduced	Reduced	Increased	Normal	Normal
2. Phosphate	Reduced	Reduced	Increased	Normal	Normal or increased
3. Bicarbonate	Normal	?	?	Increased	Increased
4. Hyperventilation	Normal	?	Normal	Normal	Increased
5. Citrate	Increased	Low	?	Normal	Normal
6. Toxic, guanidine	Increased	Normal	?	?	?
7. Magnesium deficiency	Normal	Normal	Normal	Normal	Normal

Source: After Shohl.
Tetany is due to decrease in total Ca and Ca ion (1 and 2); alkalosis (3 and 4); decrease in Ca ion (3 and 5); a toxic factor (6); and decrease in Mg (7).
* Tetany is not always present.

Calcium deficiency is by no means exceptional in human diets. It causes (*a*) disturbances in growth and development, digestive disturbances, sterility, and premature death; (*b*) osteoporosis; (*c*) slight or moderate hypocalcemia; (*d*) in some cases, hyperexcitability or even tetany.

Absorption. Calcium is absorbed from the duodenum and other parts of the small intestine. Calcium absorption is conditioned by three principal factors: (*a*) vitamin D; (*b*) the presence of other substances; (*c*) the pH of the intestinal content. Vitamin D increases absorption of Ca; it is an essential factor in the nutrition of children. Calcium salts that are soluble in water are absorbed more readily than water-insoluble calcium salts, which must first be dissolved in the intestinal content.

[1] FOOD AND NUTRITION BOARD, NATIONAL RESEARCH COUNCIL, *J. A. M. A.*, **116**, 2601, 1941.

An acid medium in the duodenum is of great importance in calcium absorption. Thus, in puppies, gastrectomy causes deficient absorption of Ca, the bones do not develop normally, and the animals become bowlegged (Ivy *et al.*). Children absorb 50 to 70 per cent of the Ca in mother's milk, but only 30 to 35 per cent of that in cow's milk. A high-protein diet favors the absorption of Ca, because calcium phosphate and carbonate are more soluble in a solution of amino acids than in water.[1] Bile and sugars (lactose) also favor calcium absorption. The absorption of fat increases that of Ca, but when fats are not absorbed, insoluble calcium soaps are formed in the intestine and Ca is thus lost in the feces. Soluble calcium salts given therapeutically should be taken outside the digestive periods.

Calcium absorption is deficient in children when the diet does not contain vitamin D or when substances are ingested that combine with Ca and form insoluble calcium salts. For example, when there is steatorrhea (fatty diarrhea), insoluble calcium salts are formed and Ca is not absorbed. When alkali, phosphate, or oxalate is ingested, Ca is not well dissolved or absorbed. The unfavorable effect of cereals on calcium metabolism (Mellanby, 1925) is due to phytic acid, which forms insoluble calcium, magnesium, and iron salts and prevents the absorption of these elements (McCance). The unfavorable effect of black bread or 85 per cent milled flour on calcium absorption should be compensated by adding 1 gm. or more of Ca per kg. of these foodstuffs.[2]

Excretion. The intestine excretes approximately two-thirds (50 to 90 per cent) of the calcium output. The remaining third is eliminated by the kidney— 30 to 50 per cent if the diet does not contain much Ca, and 10 to 20 per cent if it does. The daily excretion of Ca is usually 0.4 to 0.8 gm. in the feces and 0.15 gm. in the urine. Large amounts of Ca are secreted with the digestive juices, but most of this is reabsorbed. Renal excretion of Ca is closely related to the calcemic level. When blood calcium falls below 8.5 mg. per cent, urinary calcium diminishes, and there is little or no Ca in the urine when the blood Ca level is below 6.5 mg. per cent. Hypercalcemia increases calciuria and the proportion of total Ca eliminated by the kidney. Lactating females excrete large quantities of Ca in the milk, which should be replaced by increasing the calcium in the diet.

Calcium excretion increases considerably in hyperparathyroidism; it is high in hyperthyroidism and frequently in cases of hyperfunction of the adrenal cortex and the anterior hypophyseal lobe, as occurs in cases of acromegaly and basophil adenoma (Cushing's disease). In the latter there is osteoporosis, which involves the vertebrae and causes curvature of the spine; the bones are soft or brittle, and spontaneous fractures are observed.

Calcemia. Blood plasma contains 9 to 11.5 mg. (average 10 mg.) Ca per 100 cc. There is very little Ca in the erythrocytes (0.5 mg. per cent). Plasma Ca exists in two forms, nondiffusible and diffusible. Nondiffusible Ca (4 to 5 mg. per cent) is not separated from plasma by dialysis or ultrafiltration; it is probably bound to protein. Diffusible Ca (5 to 6.5 mg. per cent) is almost completely dissociated (ionized Ca); less than 0.25 mg. per cent is in the state of nondissociated salt. Ionized Ca concentration can be determined by its effect on the isolated frog's heart.[1] Calcium concentration in the cerebrospinal fluid is almost the same as that of diffusible Ca in plasma; it has therefore been taken as an index of diffusible and ionized Ca in blood. According to McLean and Hastings, blood proteins and Ca are in equilibrium; protein and Ca concentration rise and fall together. Ca ionization is determined by this equilibrium. At 25°C. and pH 7.35, it can be expressed as follows:

$$\frac{(Ca^{++}) \times (\text{protein})}{(\text{Ca proteinate})} = K = 10^{-2.22}$$

Therefore, knowing the protein and Ca concentration of plasma, it is possible to calculate from this equation the Ca ion concentration. A nomogram has been constructed that facilitates calculations.

Variations in calcemia. The blood-calcium level is dependent on (*a*) the amount of Ca absorbed; (*b*) vitamin D in the diet; (*c*) plasma-protein concentration; (*d*) parathyroid function. It is usually inversely related to phosphate concentration in plasma.

Hypocalcemia is observed in the following circumstances: (*a*) deficient absorption or excessive

[1] McCance, R. A., E. M. Widdowson, and H. Lehmann, *Biochem. J.*, **36**, 686, 1942.

[2] Some cereals, *e.g.*, wheat, have phytases that partially destroy phytic acid in the process of baking.

[1] McLean, F. C., and A. B. Hastings, *J. Biol. Chem.*, **107**, 337, 1934; **108**, 285, 1935.

elimination of Ca (moderate or slight hypocalcemia); (*b*) parathyroid insufficiency (marked hypocalcemia); (*c*) some cases of vitamin D deficiency; (*d*) hypoproteinemia (nephrosis) in relation to the fall in plasma-protein concentration (moderate hypocalcemia); (*e*) hyperphosphatemia in some cases of renal insufficiency (moderate hypocalcemia).

Hypocalcemia can be prevented or remedied by the ingestion of calcium; 700 cc. of milk, 11.2 gm. calcium gluconate, and 7.2 gm. calcium lactate contain 1 gm. of Ca. Intravenous injection of calcium salts (*e.g.*, 10 cc. of a 20 per cent Ca gluconate solution, injected very slowly) can be given in cases of emergency, but calcemia soon returns to the initial level. Parathyroid extract provokes a more prolonged rise in calcemia, but after a few weeks of treatment it loses its effect. Large doses of vitamin D and dihydrotachysterol are the most efficient therapeutic agents for producing a prolonged increase in the blood-calcium level and treating cases of decalcification.

Hypercalcemia is observed in the following circumstances: (*a*) immediately after intravenous injection of Ca and for 1 or 2 hours thereafter; (*b*) usually in hyperparathyroidism; (*c*) after the administration of large doses of vitamin D or of dihydrotachysterol (a substance chemically related to vitamin D); (*d*) frequently when there is hyperproteinemia (myeloma); (*e*) in certain cases of bone tumors. Calcium is taken from the bones in *b*, *c*, *d*, and *e*; also when bones are kept immobile in plaster casts.

The Regulation of Calcium Metabolism

Calcium metabolism is conditioned by the existence of a sufficient amount of Ca in the diet, its absorption in the intestine, calcium storage, blood Ca equilibrium, and calcium excretion. Certain accessory food factors, mainly vitamin D, and certain hormones, principally the parathyroid hormone, regulate different aspects of Ca metabolism.

Vitamin D. This vitamin (called the "antirachitic" vitamin) is especially important for the nutrition of the growing child. If the diet does not contain it in sufficient amounts, there are disturbances in the development of the bones and teeth, a condition known as rickets. Vitamin D increases intestinal absorption of calcium and phosphorus and normal calcification of bones and teeth. In rickets there is an insufficient absorption and retention of Ca and P. Blood calcium is usually normal or only slightly diminished, but there is a definite decrease in plasma phosphate. A fall in blood Ca, which causes tetany, is occasionally observed. Large doses of vitamin D or adequate doses of dihydrotachysterol raise the blood Ca level by increasing absorption of Ca and mobilizing bone Ca to some extent; therefore the diet must contain a sufficient amount of Ca to prevent decalcification of the bones when this treatment is applied. Very large doses of either of these steroids provoke hypercalcemia, hypercalciuria, decalcification of the bones, calcium deposits in soft tissues, and other harmful effects that may be fatal.

Vitamin C. In cases of scurvy (vitamin C deficiency), treatment with ascorbic acid (vitamin C) increases the deposition of calcium in the bones. This vitamin is not only a factor in bone calcification; it also conditions the integrity of the protein matrix of bone and the vitality of the osteoblasts.

Parathyroids. Parathyroid hormone regulates Ca metabolism, and more especially the blood-calcium level. Total parathyroidectomy does not modify the absorption of calcium, but it is followed by a decrease in blood calcium, which falls to 7 and even 5 mg. per cent, and by an increase in blood phosphate. Hypocalcemia causes a decrease in the excretion of Ca, especially of urinary calcium. As ionized Ca is also diminished, there is neuromuscular hyperexcitability, which causes spasticity, tremor, and convulsions, *i.e.*, parathyroid tetany.

Tetany is due to hypocalcemia, and all its symptoms rapidly disappear when Ca is administered and the blood-calcium level is raised.

In chronic parathyroid insufficiency there are disturbances in the calcification of the enamel and dentine of teeth and sometimes of bone, and there is deposition of Ca in the crystalline lens of the eye (cataract).

When there is an excess of parathyroid hormone, due either to repeated injections of sufficiently large doses of parathyroid extract or to excessive secretion of the hormone, hypercalcemia is observed. In cases of parathyroid adenoma, there is permanent hypersecretion of parathyroid hormone and a condition known as hyperparathyroidism. In this condition the blood calcium is high and the inorganic phosphorous low; blood phosphate is increased, and

the bones are decalcified. Large doses of parathyroid extract can raise the blood-calcium level to 18 and 20 mg. per cent and provoke signs and symptoms of acute calcium intoxication: anorexia, vomiting, general weakness, and depression, followed by coma and death if hypercalcemia is sufficiently marked and prolonged. In cases of hyperparathyroidism, the renal excretion of Ca and P is increased and sometimes calcium stones are formed in the urinary tract. Calcium and phosphorus are removed from the bones with fibrous degeneration and the formation of cysts (osteitis fibrosa cystica). Decalcification of the bones causes softening and deformation of the skeleton and curvature of the spine, and spontaneous fractures are frequently observed. The calcium requirement is increased, but even with large doses of calcium it is not always possible to obtain a positive calcium balance.

Parathyroid hormone therefore maintains the normal calcium level. Reciprocally, the calcium level regulates parathyroid secretion, which increases when there is hypocalcemia (see Chap. 55).

Thyroid. In thyroid insufficiency the growth of bones (especially in length) is arrested or proceeds at a slower rate than normal. The development of teeth is also abnormally slow. The calcium content of bones and teeth is usually lower in thyroidectomized animals than in controls of the same age. The limbs of growing subjects or animals suffering from thyroid insufficiency are short because of the disturbance in the endochondral ossification of the epiphyseal cartilages.

Hypophysis. Endochondral and periosteal ossification are markedly retarded in anterior hypophyseal insufficiency. Consequently the bones do not develop normally; growth in length is particularly disturbed. There is also dental malformation. Permanent teeth erupt later and grow more slowly than in normal animals; the pulp cavity is large owing to insufficient development of the dentine (Erausquin). In the rat, on the contrary, the pulp cavity is gradually obliterated by calcification of dentine (Schour).

In acromegaly there is excessive development of the bones of the face and cranium, especially of the jawbone, and of the hands and feet, caused by hypersecretion of the anterior hypophyseal growth hormone. In hypophyseal gigantism the bones grow abnormally long. In advanced stages of anterior hypophyseal hyperfunction, calcium excretion is increased. In acromegaly osteoporosis is frequently observed, and the patients become hunchbacked owing to curvature of the spine. In Cushing's syndrome (pituitary basophilism) there are osteoporosis, dorsocervical kyphosis, and frequently spontaneous fractures.

PHOSPHORUS METABOLISM

Functions of phosphorus. There is 450 to 700 gm. of phosphorus in the body of an adult man; 70 to 80 per cent is in the bones and teeth (mainly bound in calcium salts), and 20 to 30 per cent in other tissues. Compounds of phosphorus probably fulfill more varied functions than those of any other mineral element in the organism. Phosphoric acid is an essential constituent of nucleic acids, nucleotides, phospholipids, and hexosephosphates. It is found in many proteins, especially in the so-called "phosphoproteins" such as casein. Some of the most important enzymes are formed by vitamins combined with phosphoric acid; *e.g.*, cocarboxylase is made up of thiamine (vitamin B_1) and phosphoric acid; phosphoriboflavin is one of the respiratory enzymes; coenzymes I and II are nucleotides formed by phosphoric acid, nicotinic amide, adenine, and a pentose. Phosphorylation plays an important part in the selective absorption of sugars and fats. Glycogen is disintegrated into glucose by a process of phosphorolysis, and phosphorylation is the initial step in the synthesis of glycogen from glucose. Phospholipids are the form in which fats are most easily transported and have the highest reactivity. Adenosinephosphate and phosphocreatine in muscle set free large amounts of energy on disintegrating in the course of muscular contraction. Phosphate is the principal anion in intracellular fluids; it plays a part in the regulation of acid-base equilibrium in body fluids and in the urine.

Ingestion and requirement. The diet of a normal adult must contain a minimum of approximately 0.9 gm. of phosphorus to maintain the phosphorus balance in equilibrium; it is nevertheless advisable there should be 50 per cent more (1.35 to 1.4 gm.) as a safety margin. In pregnancy 1.5 to 3 gm. daily is needed to provide for the fetus, which accumulates 40 gm. in the course of intrauterine development. The growing child requires more P than the adult;

80 mg. per kg. is needed by the three-year-old child, 35 mg. per kg. at sixteen years, and 12 mg. per kg. by the adult.

Milk and milk products contain relatively larger quantities of phosphorus than other foods. Moreover, they also contain large amounts of calcium, whereas protein foods such as meat and fish have much phosphorus but little calcium. Meat, liver, eggs, and cereals follow in decreasing order of P content. Phosphorus in cereals is mainly bound in phytin, which is poorly absorbed (McCollum). The diet should have a phosphorus-calcium ratio of 2:1, and care should be taken that this ratio does not fall below 1:1.

Absorption. Phosphorus is mainly absorbed in the upper part of the small intestine. Soluble inorganic phosphates are more readily absorbed than the insoluble inorganic phosphates. Organic compounds of phosphorus are disintegrated by pancreatic or intestinal juice, and the P contained in them is absorbed as phosphate. Vitamin D is an important factor in the absorption of phosphorus; in vitamin D deficiency (rickets), P is absorbed in less than normal quantities. Vitamin D is a factor in the deposition of P in bone, thus decreasing its excretion; this is evident in rickets, in which not only Ca but also P are not stored in the bones and P excretion increases. Beryllium, iron, aluminum, magnesium, strontium, and an excess of Ca diminish the absorption of P, because they form insoluble compounds of phosphorus. Rickets owing to phosphorus deficiency can be produced by ingestion of these minerals.

The administration of a radioactive isotope of P (P^{35}, which has a half life of 14.3 days) has been used to demonstrate the path followed by phosphorus and its fate after it has been absorbed. There is a continuous and rapid exchange of phosphorus in the liver, kidney, and heart; phosphorus absorbed is taken up by these tissues and is retained in part for approximately 30 days. The stores of phosphorus in bone and tissues are not static; on the contrary, there is continuous intake and output of P into and from these structures.

Excretion. A slightly larger proportion of P is eliminated by the kidney than by the intestine. About 60 per cent of total P excreted is found in the urine, and an even larger percentage if stored P is mobilized, as in hyperparathyroidism. In cases of deficient absorption of P (rickets, steatorrhea, celiac disease, etc.), there is relatively more P in the feces than in the urine.

Phosphorus deficiency. Deficient absorption or utilization of P and Ca produces rickets in children and osteoporosis or osteomalacia in adults. The most frequent cause of these disturbances is vitamin D deficiency, but they may be due to insufficiency of P in the diet or to faulty absorption owing to the formation of insoluble compounds of P in the intestine. It is difficult to prepare a diet entirely free from P, but a low-P diet retards growth, the bones are poorly calcified, and the general condition is unsatisfactory; in extreme cases, the animals die Dietary phosphorus deficiency is seldom observed in man, but it is frequently observed in cattle in certain districts of South Africa or South America in which the soil and pastures have little phosphate. The animals have softening of the bones, and they eat the bones of dead animals (osteophagia). In extreme cases they also devour any solid object, such as tins, cinders, and clothes. This "perverted appetite" ceases when the phosphate content of the diet is increased.

Blood phosphorus. Blood contains inorganic phosphate and three forms of phosphorus in organic compounds: (*a*) ethereal phosphorus (glycerophosphate, hexosephosphate, etc.), which amounts to 24 mg. per 100 cc. and is found almost exclusively in the blood cells; (*b*) lipid phosphorus (phospholipids), of which there is 13 mg. in the cells and plasma; (*c*) nucleic phosphorus (in nucleic acids), of which there is only a small quantity. Total organic blood phosphorus amounts to 40 mg. per 100 cc.

There is approximately 3 to 4.5 mg. per cent inorganic phosphate in blood plasma of adults and 4 to 6 mg. per cent in children. It is diffusible phosphate and is found in the same concentration in blood plasma as in tissue fluids.

Hypophosphatemia is observed in (*a*) rickets due to vitamin D deficiency (inorganic phosphate is diminished); (*b*) osteomalacia, in which the bones soften, especially those of the pelvis (the disease is more frequent in women than in men); (*c*) hyperparathyroidism, in which there is also hypercalcemia and an increase in P excreted in the urine (in the advanced stages of the disease there are lesions in the kidney, P is retained, and the blood P may reach a normal or even a higher than normal level); (*d*) deficient absorption of P, owing to the formation of insoluble P com-

pounds in the intestine (rickets due to beryllium, strontium, etc.), or to disturbances in the intestinal mucosa (steatorrhea, celiac disease). The absorption of large quantities of glucose and insulin hypoglycemia cause a decrease in blood phosphorus, because hexosephosphate is formed and taken up by the tissues at a rapid rate.

Hyperphosphatemia is observed (*a*) when large doses of vitamin D or dihydrotachysterol are given, owing to mobilization of P stored in the bones (these substances also increase blood phosphates in hypophosphatemia of rickets); (*b*) in parathyroid insufficiency; (*c*) in severe diabetes; (*d*) in chronic renal disease, when plasma phosphate can rise to 8 mg. per cent and in extreme cases to 40 mg. per cent, as a result of retention of P (sometimes calcium phosphate is deposited in the soft tissues); (*e*) immediately after a bone is fractured, when blood phosphate usually increases for a short time.

Phosphatases. Enzymes that accelerate hydrolysis of monophosphoric esters, releasing phosphoric acid, are known as phosphatases.[1] They are found in nearly all the cells and tissue fluids. There are two phosphatases in blood plasma. One exerts maximum activity at pH 9; this is called alkaline phosphatase. There are large quantities of alkaline phosphatase in the intestine (absorption), kidney (excretion), liver (formation of phosphate compounds), and bones and teeth (ossification). The principal source of alkaline phosphatase appears to be the skeleton. It is increased in the blood in cases of rickets and other bone diseases (*e.g.*, Paget's disease), and less markedly in hyperparathyroidism, jaundice, and other disturbances of the liver, etc.

Acid phosphatase has its maximum activity in an acid medium. It is found in the liver, kidney, erythrocytes, etc. It is found in large quantities in the prostate, and in cases of cancer of the prostate it increases in the blood.

BONE AND THE FORMATION OF BONE

Composition. Bone in the body is a living tissue, made up of an organic matrix (30 per cent) impregnated with inorganic salts (45 per cent) and containing 25 per cent water. The inorganic salts in bone can be dissolved and removed by treatment with acid; the organic matrix that remains is elastic and flexible. This matrix is formed chiefly by a collagenous protein, called ossein. Calcination of bone destroys the organic matrix, and there remains a brittle structure (bone ash) which represents 45 per cent of fresh bone and 60 per cent of dried. This ash contains 37 per cent of calcium, 17 per cent of phosphorus, 0.5 per cent of magnesium, and small quantities of potassium, sodium, iron, chloride, and fluoride; the calcium-phosphorus ratio is 2:1. Calcium phosphate makes up about 85 per cent of the inorganic constituents of the bone, and calcium carbonate 12 per cent. The study of the crystalline structure of bone by means of x-ray diffraction has shown it to consist largely of a substance similar to apatite or hydroxyapatite, $Ca_{10}(PO_4)_6(OH)_2$.

Functions of bone. The skeleton is the supporting framework of the body. It owes its rigidity to its mineral salts, and its elasticity to the organic matrix. When bone loses its minerals (decalcification) it loses its rigidity, softens, and is bent (rickets, osteomalacia, hyperparathyroidism). When the proportion of the organic components diminishes, bone becomes brittle and fractures easily (marble or chalky bone, or Albers-Schönberg disease).

There is active interchange of calcium and phosphate between bone tissue and the tissue fluids and plasma. When radioactive phosphorus (P^{35}) is administered, it is found in the bones after a few hours. Phosphorus and calcium stored in the bones are mobilized whenever there is deficiency of Ca. Thus when the diet does not contain a sufficient amount of Ca the blood Ca level does not fall; it is maintained by removing Ca from the bones, which are gradually decalcified (osteoporosis). Marked decalcification is also provoked by excess parathyroid hormone. Faulty deposition of Ca in bones is observed when there is vitamin D deficiency (rickets), and in thyroid and anterior hypophyseal insufficiencies.

Other elements besides Ca and P are deposited in the bone, *e.g.*, lead, radium, arsenic, and fluorine. If one of these substances stored in the bone (*e.g.*, lead) is suddenly mobilized, toxic symptoms (*e.g.*, lead poisoning) may be produced. For this reason a detoxicating function has been attributed to bone tissue. Substances

[1] Kay, H. D., *Physiol. Rev.*, **12**, 384, 1932; Bodansky, A., *J. Biol. Chem.*, **101**, 93, 1933; **104**, 473, 1934; Sunderman, F. W., *Am. J. Clin. Path*, **12**, 404, 1942; Jaffe, H. L., and A. Bodansky, *Bull. New York Acad. Med.*, **19**, 831, 1943.

stored in the bone can, nevertheless, produce harmful effects. Thus radium deposited in bones destroys the hemopoietic tissue in the bone marrow and may cause fatal aplastic anemia. An excess of fluoride in the bones alters their structure; they become hard and brittle (osteopetrosis) and bony excrescences (exostoses) are formed, but in some cases there are also decalcified areas and cysts.[1]

Ossification. Bone is formed from cartilage (endochondral ossification), or by the osteogenic cells under the periosteum (periosteal ossification). As soon as bone is formed it undergoes a process of reabsorption and replacement, which is very active in the fetus and child but diminishes as age advances. Bone increases in thickness as new layers are laid down under the periosteum; simultaneously bone is reabsorbed, and the central medullary canal is enlarged to such an extent that the femur of a fetus can be introduced into the central canal of an adult's femur. This process can be demonstrated by adding alizarin to the food. The dye is deposited in newly formed bone and gives it a red color. The new red layer is found under the periosteum if the bone is examined soon after the dye has been given. As time passes, it is displaced toward the central cavity by the more recent layers, until it forms the innermost layer around the central canal and is finally reabsorbed.

Growth in length is achieved by ossification of the epiphyseal cartilage (endochondral ossification). Cartilage cells proliferate and are disposed in columns; later they become swollen and vacuolized. Calcium salts are deposited in the organic matrix that surrounds these cells. Blood vessels and cells from the bone marrow then penetrate the ossifying cartilage, and the final structure of bone is completed. Endochondral ossification cannot proceed normally if there is deficiency in vitamin D or phosphorus, as in rickets. The cartilage cells continue to proliferate, the epiphyseal cartilage widens, and the cell columns become irregular, but little calcium is deposited in the osteoid layer, and ossification is not completed. In thyroid and hypophyseal insufficiencies the development of the cartilage cells and ulterior ossification are retarded. In normal subjects ossification is complete and the epiphyseal plate is closed at the age of twenty or twenty-five years. In castrates epiphyseal endochondral ossification continues for several years more; for this reason their limbs are abnormally long. Excess secretion of anterior hypophyseal growth hormone provokes hypertrophy and hyperplasia of the cartilage cells and causes giantism or acromegaly.

Parathyroid hormone regulates the reabsorption of bone. An excess of this hormone produces decalcification of bone and the disease known as osteitis fibrosa cystica.

The process by which crystallized calcium phosphate is deposited in the bones is not yet completely known. An organic matrix is formed, and then Ca phosphate is deposited; later the crystal is formed by the addition of Ca ions, hydroxyl, and calcium carbonate. Finally most of the calcium is in crystals similar to apatite (carbonate apatite).

Saturation and precipitation in a calcium phosphate solution are determined not by the total amounts of calcium and phosphate, but by the product of the concentration of calcium and phosphate ions, *i.e.*, $[Ca^{++}] \times [PO_4^{\equiv}]$. The ion product at which the solution is just saturated is known as the "solubility product." In blood plasma part of the Ca is bound in Ca proteinate, which is a weak electrolyte and therefore has a low degree of dissociation. Plasma thus dissolves more calcium than is dissolved in distilled water. CO_2 also increases calcium solubility. A decrease in plasma protein or in CO_2 diminishes the solubility of calcium which tends to precipitate. Phosphate concentration in plasma seems to be of greater importance than the solubility product for the process of bone formation, because the administration of phosphate is more efficacious than that of Ca in the treatment of rickets in which plasma-phosphate concentration is below normal.

According to Robison,[1] phosphatase plays an important part in ossification. This enzyme disintegrates monophosphate esters (hexosephosphate, glycerophosphate) and releases PO_4 ion, thus increasing the ion product $[Ca^{++}] \times [PO_4^{\equiv}]$ locally where it acts, *e.g.*, in bone cartilage. When the ion product exceeds the solubility product, *i.e.*, the solution is supersaturated, calcium phosphate precipitates and is deposited in

[1] A certain amount of fluorine strengthens the dentine in teeth, and increases the resistance to caries, but a larger percentage of fluorine in the teeth "mottles" the enamel. Large doses of fluorine cause disturbances in bones and teeth (see "Fluorine," page 491).

[1] Robison, R., and K. M. Soames, *Biochem. J.*, **18**, 740, 1924.

the organic matrix of bone cartilage. The following facts support this theory: (*a*) phosphatase is in high concentration in bones and teeth, especially in the epiphyseal cartilage when active calcification is taking place, but there is none in resting or nonossifying cartilage, such as that on the surface of joints; (*b*) rachitic bones contain phosphatase, so when they are submerged in a solution of calcium hexosemonophosphate, calcium phosphate precipitates. Certain facts, however, are not explained by this theory; thus, there are tissues which do not ossify but in which there are large quantities of phosphate. Moreover there is very little hexosemonophosphate in tissue fluids. Robison has postulated the existence of two different mechanisms that take part in ossification. One is the phosphatase mechanism; the other is not yet well known but is inhibited by cyanide.

The organic matrix seems to contain a factor of ossification that is resistant to heat, because a piece of boiled bone cartilage ossifies if it is placed in the peritoneal cavity, although its cells and enzymes have been destroyed, also, bone is formed if an alcoholic extract of epiphyseal cartilage is injected into a muscle.[1]

The formation of bone is the result of an equilibrium between local processes and the composition of plasma and tissue fluids, which regulates the continuous deposition of calcium salts in the bones and their reabsorption. The following factors are known to play a part in this process: (*a*) the balance between the calcium intake and excretion; (*b*) vitamin D; (*c*) plasma calcium and phosphate concentration; (*d*) hormones secreted by the parathyroid, anterior hypophysis, thyroid, and gonads; (*e*) vitamin C, which has some influence on the vitality of osteoblasts and odontoblasts; (*f*) accidental factors, such as fluoride and perhaps others still unknown, which regulate the formation of ossein.

IRON METABOLISM

Functions and distribution. Iron is essential for life. It is necessary for the formation of erythrocytes (hemopoiesis). It forms part of respiratory mechanisms such as blood hemoglobin, myohemoglobin, and cytochromes. There is 5 gm. of iron in the body; 65 per cent is in hemoglobin; 15 per cent in the liver, spleen, and bone marrow; and 20 per cent in other tissues. Hemoglobin has 0.34 per cent iron; therefore in 100 cc. of blood there is 52 mg. of iron in man and 45 mg. in women. There is very little iron in blood plasma (0.08 to 0.12 mg. per cent).

Requirement. The minimum iron requirement is probably 5 mg. per day, but a normal diet should contain from 10 to 20 mg. daily. An adult man or woman should receive 12 mg.; a pregnant or lactating woman, 15 mg.; a child during the first year of life, 6 mg. The intake should rise from 7 to 10 mg. between the second and tenth year, and up to 12 mg. by the twelfth year.[1]

Relatively large quantities of iron are found in liver, meat, egg yolk, green vegetables, and legumes. Milk contains very little iron. If it is the sole food for a long time, the subject becomes anemic when the iron stores are exhausted. Human milk has more iron than cow's milk. Repeated hemorrhages, even if they are small, *e.g.*, intestinal hemorrhage caused by hookworm, provoke anemia due to iron deficiency; a meat diet or iron treatment (O. Cruz) prevents or cures this type of anemia. Anemia due to iron deficiency is of the microcytic hypochromic type; it is more frequent in women and children than in men. Total deprivation of iron has been observed only in laboratory experiments. It is fatal if sufficiently prolonged.

Iron is absorbed in the duodenum, probably as a ferrous salt; after it is absorbed it is carried in the plasma as a ferric salt. Experiments with radioactive iron, which has a half life of 47 days, have shown that iron absorption is regulated by a mechanism in the intestinal mucosa, so that the intake is adjusted to the needs of the organism. Normally very little iron is absorbed, but as soon as the store of iron in the body diminishes, the amount absorbed increases.[2] The acidity of the gastric juice is an important factor in iron absorption, which diminishes when there is anachlorhydria. Iron absorption diminishes when substances are ingested that form insoluble iron salts, *e.g.*, phytic acid, brown bread, etc. In wartime, when there is a large consumption of brown bread, iron should be added to the flour.

Iron is eliminated by the cecum and the large

[1] LEVANDER, G., *et al.*, *Nature*, **155**, 148, 1945; **157**, 587, 1946.

[1] FOOD AND NUTRITION BOARD, NATIONAL RESEARCH COUNCIL, *loc. cit.*

[2] HAHN, *et al.*, *J. Exper. Med.*, **69**, 739, 1939; **78**, 169, 1943.

intestine. Most of the iron in the feces comes from food that has not been absorbed; only a small part is iron excreted from the body, because although hemoglobin is being continuously destroyed, its iron is retained and used in the formation of new hemoglobin.

A crystallized protein, called "ferritin," has been obtained from the liver and spleen; it has 20 per cent iron.[1] It apparently plays a part in intestinal absorption and in storage of iron. It has an antidiuretic effect and seems to be the vasodepressor substance released by the liver during shock (Mazur and Shorr). It controls the passage of iron through the intestine, kidney, and placenta. It combines with a protein (transferrin) which carries iron in the blood plasma.

The following disturbances in iron metabolism may be mentioned: (*a*) there may be hypochromic anemia due to dietary iron deficiency; (*b*) in hemosiderosis a brown pigment called hemosiderin, containing iron, accumulates in the reticuloendothelial system; (*c*) after splenectomy iron is not stored as efficiently as in normal subjects, and the excretion of iron increases, resulting in an iron deficiency which may cause anemia.

OTHER MINERAL CONSTITUENTS

Magnesium. This is an essential element in several enzymatic systems, such as those of phosphorylase, carboxylase, and cozymase. In plants it is found in chlorophyll and is comparable to iron in hemoglobin. Rats fed on a diet completely free from magnesium suffer from convulsions and tetany and eventually die.[2] Magnesium deficiency provokes tetany in cows that have recently calved and in lactating cows. The condition is cured by the ingestion of Mg or Ca. The minimum daily requirement of magnesium is 0.22 gm.,[3] but a normal diet should contain 0.27 gm. (Sherman).

Magnesium is absorbed in the small intestine. Phytates in cereals form insoluble Mg compounds and decrease its absorption. Magnesium is excreted mainly in the feces (50 to 80 per cent). Blood plasma has a Mg concentration of 1.8 to 3.6 (average 2.5) mg. per cent, and erythrocytes 5.4 to 7.8 mg. per cent. When the concentration in plasma rises to 5 mg. per cent, there is somnolence. Lethargy increases and the subjects become comatose as the concentration rises to 15 and 20 mg. per cent. Intravenous injection of Ca diminishes the toxic effects of magnesium.

Iodine. The daily requirement is between 100 to 200 μg (2 to 4 μg per kg.). In districts where endemic goiter is prevalent these quantities should be doubled. Iodine metabolism will be fully considered in Chap. 53, The Thyroid Gland.

Fluorine. Very small amounts of fluorine are found in the bones and teeth. If the drinking water contains 1 to 2 mg. F per liter, white patches appear on the surface of the teeth, and brown or black patches if there is 2.5 mg. or more per liter. This is due to a disturbance in the structure of enamel, which becomes porous and pigmented (mottled enamel). Sensitiveness to fluorine varies from one individual to another, but if there is 5 to 6 mg. per liter of F in the drinking water, 100 per cent of the population have mottled enamel. The structural change in the teeth takes place in infancy when the teeth are developing. If drinking water contains 1 to 1.5 mg./liter, the incidence of caries diminishes and the resistance of dentine increases. Cases of osteopetrosis (compact "marble" bone) have been reported in districts in which there is a high concentration of F in drinking water (*e.g.*, La Pampa, Argentina). Most of the patients were over thirty years old. Large doses given experimentally to rats produce severe lesions in bones and teeth. The bone is compact and petrous in patches, while in other parts there may be osteoporosis, cyst formation, or bony excrescences (osteophytes).

Zinc. Zn is apparently an indispensable element. It forms part of the carbonic anhydrase enzymatic system; it has been found in crystallized insulin and in some snake venoms. It is taken up mainly by the liver and pancreas and is excreted principally in the pancreatic juice.

Cobalt. This metal is indispensable in animal nutrition. Cobalt has been found in vitamin B_{12}. Pancreatic tissue contains cobalt and nickel (Bertrand). Cobalt deficiency has been reported in certain districts of Scotland and New Zealand, in which diseases and mortality in sheep due to this deficiency have been observed.

[1] Laufberger, M., *Bull. Soc. chim. biol.*, **19,** 1575, 1937; Granick, S., *J. Biol. Chem.*, **146,** 451, 1942; **147,** 91, 1943; **148,** 463, 1943; **149,** 157, 1943; *Bull. New York Acad. Med.*, **30,** 84, 1954; Michaelis, L., *Advances in Protein Chemistry*, **3,** 53, 1947.

[2] Kruse, H. D., E. R. Orent, and E. V. McCollum, *J. Biol. Chem.*, **100,** 603, 1933.

[3] Tibbets, D. M., and J. C. Aub, *J. Clin. Investigation*, **16,** 491, 1937.

Other elements. Manganese is indispensable to rats and chicks; it forms part of several enzymatic systems. It is eliminated mainly in the bile. Copper reinforces the effect of iron in the formation of erythrocytes and hemoglobin in rats kept on a exclusively milk diet. This effect has not been observed in man. Deficiency of Cu has never been reported; apparently the small amount of Cu in food and water is sufficient for the needs of human nutrition, which have been estimated to be 50 μg per kg. daily.

Toxic effects. Minerals sometimes produce toxic effects. Fluorine intoxication has been observed in districts with a high F content in the drinking water and near aluminum factories where F expelled in smoke is condensed by fog. Chronic intoxication by arsenic has been reported to be endemic in certain districts (Córdoba, Argentina); it produces pigmentation of the skin, hyperkeratosis, and several visceral disturbances (Ayerza). Intoxication of cattle by selenium, molybdenum, lead, or fluorine in water or pastures has been reported. This contrasts with diseases due to deficiency of some element, such as those caused by Ca, Mg, P, or Co deficiency. Intoxication or diseases in workers due to industrial utilization of minerals are fully considered in treatises on industrial hygiene.

BIBLIOGRAPHY

Baudisch, O., The Importance of Trace Elements in Biologic Activity, *J. A. M. A.*, **123**, 359, 1943.

Bland, J. H., "The Clinical Use of Fluid and Electrolytes," Saunders, Philadelphia, 1952.

Curtis, G. M., and M. B. Fertman, *J. A. M. A.*, **139**, 28, 1949.

Dallemagne, M. J., L'os et les mécanismes de sa formation, *J. de Physiol.*, **43**, 425 and 627, 1951.

Darby, W. J., Iron and Copper, *J. A. M. A.*, **142**, 1288, 1951.

Darrow, D. C., Disturbances in Electrolyte Metabolism in Man and Their Management, *Bull. New York Acad. Med.*, **24**, 147, 1948.

Deane, N., and H. W. Smith, The Distribution of Sodium and Potassium in Man, *J. Clin. Investigation*, **31**, 197, 1952.

Edelman, J. S., *et. al.*, Body Composition: Studies in the Human Being by the Dilution Principle, *Science*, **115**, 447, 1952.

Fenn, W. D., The Role of Potassium in Physiological Processes, *Physiol. Rev.*, **20**, 377, 1940.

Finch, C. A., J. A. Wolff, and C. E. Rath, Iron Metabolism, *J. Lab. & Clin. Med.*, **34**, 1480, 1949.

Fowler, W. M., and A. P. Berer, Plasma Iron, *Am. J. Med. Sc.*, **223**, 633, 1952.

Glass, B., A Symposium on Copper Metabolism, Contr. No. 5, McCollum-Pratt Institute, Johns Hopkins University Press, Baltimore, 1950.

Heat, C. W., Iron in Nutrition, in "Handbook of Nutrition," American Medical Association, Chicago, 1943.

Hemmeler, G., "Métabolisme du fer," Masson et Cie, Paris, 1951.

Higgins, C., The Composition of Bone and the Function of the Bone Cell, *Physiol. Rev.*, **17**, 119, 1937.

Hoffmann, W. S., Clinical Pathology of Potassium, *J. A. M. A.*, **144**, 1157, 1950.

Howard, J. E., Metabolism of Calcium and Phosphorus in Bone, *Bull. New York Acad. Med.*, **27**, 24, 1951.

Logan, M. A., Recent Advances in the Chemistry of Calcification, *Physiol. Rev.*, **20**, 522, 1940.

Marston, H. R., Cobalt, Copper, and Molybdenum in the Nutrition of Animals and Plants, *Physiol. Rev.*, **32**, 66, 1952.

McCollum, E. V., E. Orent-Keiles, and H. G. Day, "The Newer Knowledge of Nutrition," 5th ed., Macmillan, New York, 1941.

McLean, F. C., Physiology of Bone, *Ann. Rev. Physiol.*, **5**, 79, 1943; *Proc. Inst. Med. Chicago*, **16**, 414, 1947.

Newburg, L. H., "Significance of the Body Fluids in Clinical Medicine," Charles C Thomas, Springfield, Ill., 1951.

Peters, J. P., The Role of Sodium in the Production of Edema, *New England J. Med.*, **239**, 353, 1948.

Pitts, R. F., Modern Concepts of Acid-Base Regulation, *Arch. Int. Med.*, **89**, 864, 1952

Reifenstein, E. C., F. Albright, and S. L. Wells, The Accumulation, Interpretation and Presentation of Data Pertaining to Metabolic Balances, Notably Those of Calcium, Phosphorus, and Nitrogen, *J. Clin. Endocrinol*, **5**, 366, 1945.

Rosenheim, M. L., Sodium, *Lancet*, **261**, 505, 1951.

Schmidt, C., and D. M. Greenberg, Occurrence, Transport and Regulation of Calcium, Magnesium and Phosphorus in the Animal Organism, *Physiol. Rev.*, **15**, 297, 1935.

Sheppard, C. W., New Developments in Potassium and Cell Physiology, *Science*, **114**, 85, 1951.

Sherman, N. C., "Chemistry of Food and Nutrition," 7th ed., Macmillan, New York, 1946.

Shohl, A. T., "Mineral Metabolism," Reinhold, New York, 1939.

Stearns, G., Requirements of Calcium, Phosphorus and Magnesium, *J. A. M. A.*, **142**, 478, 1950.

Von Oettingen, W. F., Manganese: Its Distribution, Pharmacology and Health Hazards, *Physiol. Rev.*, **15**, 175, 1935.

Widdowson, E. M., R. A. McCance, and C. M. Spray, The Chemical Composition of the Human Body. *Clin. Science*, **10**, 113, 1951.

CHAPTER 46

Metabolism in Fasting, Malnutrition, and Obesity

METABOLISM IN FASTING

During a complete fast, the body functions at the expense of substances stored and taken from the tissues. Carbohydrate stores amount to a few hundred grams and are depleted in one or two days; energy is thereafter obtained mainly (80 to 90 per cent) from fat and in a lesser degree (10 to 20 per cent) from protein. In an advanced stage, when all storage fat has been burned, there is a profound change in metabolism and large quantities of protein are consumed.

Glycogen stores diminish rapidly, but even after a long fast there is glycogen in the heart and muscles and a small amount in the liver, which is formed by glyconeogenesis. The blood sugar falls a little, but it is maintained within the normal range. After a few days of fasting the capacity to utilize carbohydrate diminishes. Thus if glucose or carbohydrate is administered the rise in blood sugar is more marked and prolonged than in normal subjects; there is little or no increase in the respiratory quotient, and glycosuria is easily produced.[1]

Storage fat is the principal source of energy in fasting (Terroine's "élément variable," Chap. 42). If the fast is sufficiently prolonged 97 per cent is consumed (Voit). Tissue fat, mainly phospholipids (Terroine's "élément constant"), does not diminish much. Mobilization of storage fat causes a rise in lipemia and liver fat. The considerable increase in fat consumption leads to the production of large quantities of acetoacetic and β-hydroxybutyric acids in excess of that which can be oxidized by the organism. These acids are accumulated in the blood and excreted in the urine. Ketosis (*i.e.*, increased production and elimination of ketone bodies) causes acidosis, ketonuria, increased urinary ammonia, diminished alkali reserve, hyperventilation, and a low alveolar CO_2 tension.

Protein catabolism diminishes rapidly in the first two or three days of fasting, and therefore so does the excretion of nitrogen, sulfur, and phosphorus in the urine; later it continues to diminish gradually. Normally fed subjects eliminate 12 to 16 gm. of N daily; in subjects who had fasted for 31 to 38 days, the daily N elimination dropped to between 6.9 and 3.3 gm. The total amount of urine also decreased. Urea was the principal nitrogenous substance in the urine of these subjects; therefore urea is not only exogenous but also endogenous, *i.e.*, it is an end product of the catabolism of tissue protein. As the subjects lost weight creatinine excretion diminished, and creatine was found in the urine. Amino-acid concentration in plasma is not diminished in fasting subjects; therefore intestinal absorption is not the only source of the amino acids in the blood.

Feces are eliminated even when no food is taken. Succi, a subject studied by Luciani, continued to eliminate 9.5 to 22 gm. of feces with 0.3 to 1 gm. of nitrogen daily.

After the fat stores have been depleted, large quantities of protein are burned. Urinary nitrogen rises, and if the subject is not fed he dies. At this stage the capacity to burn carbohydrate improves; the administration of glucose is followed by a rise in the RQ (Hédon), not only in

[1] Claude Bernard first described these effects of fasting. Hofmeister (1890) gave the name of "hunger diabetes" to this condition. CHAMBERS, W. H., Undernutrition and Carbohydrate Metabolism, *Physiol. Rev.*, **18**, 248, 1938; LUNDBAEK, K., *Yale J. Biol. & Med.*, **20**, 533, 1948.

normal animals but also in diabetic ones, and the latter retain part of the glucose without excreting it into the urine (Chambers).

The mineral content of plasma is not changed, because the body retains salts, and the excretion of Cl, Na, K, and Ca diminishes. In the first stages NaCl is lost, together with extracellular water, but later NaCl excretion is reduced to a minimum. The loss of intracellular water is accompanied by the loss of potassium.

Table 59. Percentage of Weight Lost by Tissues during Starvation

Tissue	*Loss of fresh tissue, %*	*Loss of dry tissue, %*
Adipose	97	
Spleen	67	63
Liver	54	57
Testes	40	
Muscle	31	30
Blood	27	18
Kidney	26	21
Skin and hair	21	
Intestine	18	
Lung	18	19
Pancreas	17	
Heart	3	
Brain and spinal cord	3	0

Source: according to Voit.

The BMR decreases gradually; in some cases it falls by 30 and even 50 per cent. Body temperature, especially skin temperature, also drops a little. The pulse rate diminishes and intestinal absorption is slowed down. During the first days of fasting, intense hunger is felt, which diminishes later and is no longer felt after the first week. There is progressive weakness, leading to physical and mental lethargy and finally to coma and death. Occasionally hypoglycemic convulsions are observed.

Complete fasting leads to death after a time that varies in different species. Smaller animals with a higher BMR die sooner than larger animals with a low BMR. Professional fasters, such as Succi and Cetti, have fasted for 30 and even 50 days, drinking only water during that time. MacSwiney, the mayor of Cork, who refused food when imprisoned, died in coma after 74 days. A dog has been kept fasting for 117 days, without dying, in spite of having lost 63 per cent of its body weight. This is the longest period of fasting recorded in a homoiothermic animal. Poikilothermic animals survive many months of complete fasting if the environmental temperature is kept low.

Body weight is lost gradually and continuously during fasting. The decrease varies between 20 and 60 per cent, but is usually 40 to 50 per cent before the animal dies. Weight lost by the different tissues varies considerably (Table 59). Adipose tissue is almost completely lost, and about one-third of the muscle tissue is consumed.

Table 60. Bodily Changes in a Fasting Human Subject

	Day of fasting			
	First	*Eleventh*	*Twenty-first*	*Thirty-first*
Body weight, kg	59.60	53.88	50.49	47.39
Rectal temperature	......	36.54	36.04	35.96
Pulse rate	74	61	59	60
Hemoglobin, %	90	85	88	92
Alveolar CO_2 tension, mm. Hg	32.8	......	28.7	31.8*
Urine				
Total nitrogen	7.10	10.25	7.93	6.94
Urea nitrogen	5.68	7.66	5.54	4.84
Creatine + creatinine N	0.48	0.49	0.38	0.32
Chlorine	3.77	0.36	0.18	0.13
P_2O_5	1.66	1.95	1.60	1.32
S	0.46	0.62	0.51	0.49
β-Hydroxybutyric acid	......	1.4	5.0	4.5
Sodium	2.070	0.100	0.066	0.053
O_2, night, cc. per min	212	176	154	160
Cal. per sq. m., 24 hr.	843	732	653	701†

Source: BENEDICT, F. G., "A Study of Prolonged Fasting," Carnegie Institution, Publ. No. 203, Washington, D.C., 1915.

* Previous day, 27.08.

† Previous day, 698.

The spleen, liver, and testes are particularly sensitive to fasting. On the other hand the heart and central nervous system lose very little weight. Some of the bodily changes observed in a subject who fasted for 31 days are given in Table 60.[1]

[1] The older literature on fasting is well summarized by E. Bardier in Richet, "Dictionnaire de physiologie," **9**, 58, Librairie Félix Alcan, Paris, 1913, and by G. Lusk (The Physiological Effects of Undernutrition, *Physiol. Rev.*, **1**, 523, 1921).

MALNUTRITION

Malnutrition occurs when the diet is deficient in one or more of its essential components, *i.e.*, calories, protein, certain amino acids, certain fatty acids, vitamins, or minerals. Deficiency in water or oxygen can also produce conditions similar to those of malnutrition. Complete absence of an essential constituent may frequently occur in times of war or famine, but partial deficiencies sometimes remain hidden and cause widespread damage. Production, transport, and distribution of food have improved continuously since the eighteenth century, and in more recent years the knowledge of nutrition has progressed considerably, so that generally speaking human beings are now better fed than ever before. On the other hand, war has created conditions of famine in certain populations such as have never been seen on so wide a scale.

Table 61. Effects of Malnutrition Due to Dietary Deficiency, Increased Metabolism, or Deficient Absorption

Cause of malnutrition	*Effects*
Deficiency in	
Calories	Loss of weight, weakness
Protein	Hypoproteinemia, edema, microcytic anemia, muscular wasting
Vitamin A	Keratosis in conjunctiva and mucosae; night blindness
Vitamin B_1 (thiamine)	Mental and physical fatigue, anorexia, polyneuritis, muscular cramps and pains, cardiac dilatation
Nicotinic amide	Pellagra
Riboflavin (vitamin B_2)	Weakness, cheilosis, stomatitis, nasolabial seborrhea, interstitial keratitis
Vitamin B complex	Deficiency in one of the vitamins is usually accompanied by some degree of deficiency in the other vitamins of the complex
Ascorbic acid (vitamin C)	Scurvy, capillary fragility
Vitamin D	Rickets
Tocopherol (vitamin E)	Sterility in rats (not proved in man)
Vitamin K	Prolonged blood-clotting time, diminished prothrombin
Calcium	Rickets in children; osteoporosis and osteomalacia in adults, especially in multiparous and lactating women
Iodine	Endemic goiter
Iron	Hypochromic anemia, especially in children and pregnant women and after repeated hemorrhage. Hookworm anemia
Chronic starvation	The effects of several deficiencies are observed
Increase in metabolism owing to	
Growth	Rickets, osteoporosis, anemia, retarded growth
Pregnancy, lactation	Child: rickets, anemia, retarded growth Mother: anemia, osteoporosis, weakness
Thyrotoxicosis, fever, leukemia	Loss of weight if additional calories to cover increased metabolism are not taken
Deficient absorption due to	
Deficiency in gastric acidity	Disturbances in bone
Deficiency in pancreatic lipase owing to (1) insufficient production (chronic pancreatitis) (2) obstruction of duct	Insufficient absorption of fat and vitamin A; steatorrhea; chronic undernutrition
Chronic diarrhea	Loss of fat, protein, vitamins, Ca, Fe; anemia, osteoporosis, tetany, edema, polyneuritis, pellagra, undernutrition

Source: after Newburg, with modifications.

Malnutrition may be caused by (*a*) insufficient ingestion of food; (*b*) deficient absorption of food; (*c*) increase in metabolism. Insufficient ingestion of food is frequently due to anorexia (lack of appetite), to food fads that cause the exclusion of an essential component or its reduction to quantities below the safety level, or to limitation of the diet to a few chosen foods, *e.g.*, white bread, sugar, and margarine. A frequent

cause of malnutrition in children and pregnant or nursing women, in many regions, is the scarcity of milk and milk products. Deficiencies in Ca, vitamin D, animal protein, vitamin B complex, and vitamin C are among those most commonly observed. Table 61 gives a summary of the outstanding symptoms typical of the most frequently found deficiencies.[1]

The signs and symptoms of malnutrition are similar to those of starvation. They have been observed in countries at war, in times of famine, in concentration camps, and also in experimental studies made on volunteers.[2] There is loss of weight and storage fat, which may be masked by edema. The subjects may become extraordinarily thin. The BMR falls 15 to 32 per cent and the body temperature diminishes slightly (0.5 to 0.9°C.). There is bradycardia, the heart is smaller, and the blood pressure decreases (about 20 per cent). The erythrocytes and hemoglobin are slightly below normal. The subjects are physically weak and mentally exhausted; they are apathetic and apparently lazy. Susceptibility to infections increases. Acidosis occurs frequently, especially in fat subjects, but it is usually well tolerated. Intense and prolonged undernourishment may cause hypophyseal and sexual functions to diminish.

In certain forms of malnutrition prevalent in parts of Africa and Central and South America, there is a high infantile morbidity and mortality. Children suffering the disease known in Africa as *kwashiorkor* (red child), due mainly to protein deficiency,[3] have red or straw-colored hair, staining or thickening of the skin, edema, anemia, gastrointestinal disturbances, and fatty degeneration or cirrhosis of the liver; the mortality is high (Gillman).

A well-balanced and plentiful diet improves growth, raises the average height, and increases the average life span of any population.[4]

[1] KRUSE, H. D., *et al.*, National Research Council Bulletin No. 109, Nat. Acad. Sciences, Washington, D.C., 1943; *Bull. Pan-American Sanit. Bureau*, **21**, 564, 1942; KRUSE, H. D., *J. A. M. A.*, **121**, 584, 1943.

[2] BENEDICT, F. C., Carnegie Institution, Pub. No. 280, Washington, D.C., 1919.

[3] DEAN, R. F. A., H. C. TROWELL, and J. N. P. DAVIES, *Brit. M. J.*, **2**, 791, 1952.

[4] In countries where food is plentiful, such as the United States and Argentina, the offspring of European immigrants are on an average taller than their parents, who were brought up in a less well-fed community Mendel obtained an increase in size in white rats by improving the diet.

OBESITY

Obesity is an important problem in nutrition and preventive medicine. More than one-fifth of the adults in well-fed countries are obese. Degenerative diseases are more frequently observed in fat than in thin persons; also the average life is shorter in fat people.

Usually 15 to 17 per cent of the body weight is fat, but this varies considerably in different individuals. The fat content can be determined in living subjects by (*a*) measuring the body water (McCance and Widdowson); (*b*) measuring the specific weight of the body (Behnke), since fat is less heavy, having a specific weight of 0.92; (*c*) measuring the thickness of skin and subcutaneous folds; (*d*) anthropometric measurements and x-rays.

Obesity is sometimes classified, according to the mechanism of its production, into (*a*) alimentary or exogenous obesity, due to excessive food intake; (*b*) constitutional or endogenous obesity, attributed to the hereditary constitution of the subject; (*c*) endocrine obesity, which may be due to disturbances in the functions of the hypophysis (dystrophia adiposogenitalis; Cushing's disease), adrenals (hypercorticalism), thyroid (hypothyroidism), or gonads (menopause; castration); (*d*) obesity due to brain lesions, especially those of the hypothalamus. The diet is a fundamental factor in all forms of obesity, since fat stored has its origin in food. Obesity diminishes when food is restricted; there were no obese persons in Europe at the end of the Second World War. Other factors contribute to increase fat storage by different mechanisms. Endocrine factors will be considered in the chapters on the respective glands. Obesity due to brain lesions is usually preceded by an increase in appetite (hyperphagia); this has been observed in animals with lesions in the hypothalamus[1] and after removing the frontal lobe.[2]

The metabolism of obese subjects has been carefully studied.[3] Usually there is no abnormality in the BMR. The specific dynamic effect is normal or slightly below normal, but this decrease does not cause the total metabolism to

[1] HETHERINGTON, A. W., and S. W. RANSON, *Am. J. Physiol.*, **136**, 609, 1942; BROBECK, I. R., *et al.*, *Yale J Biol. & Med.*, **15**, 831, 855, and 893, 1943.

[2] BEACH, F. A., *J. Comb. Psychol.*, **31**, 145, 1941.

[3] EVANS, F. A., in Duncan, "Diseases of Metabolism," Saunders, Philadelphia, 1947; NEWBURGH L. J., and J. W. CONN, *Physiol. Rev.*, **24**, 18, 1944.

fall by more than 3 per cent. Nitrogen metabolism is normal and nitrogen equilibrium is maintained with 1 gm. of protein per kg. daily even if the total caloric intake is reduced to 400 kg.-cal. Creatinine excretion is proportional to the muscular development of the subject; the daily elimination per kilogram of total body weight is below normal.

Familial factors play a part in some cases of obesity, *e.g.*, a constitutional deviation in metabolism. For instance, lipophilia, *i.e.*, abnormally high formation and storage with low utilization of fat, is often supposed to be a cause of obesity, but this has never been satisfactorily demonstrated. Familial customs, such as overeating and sedentary habits, are frequently the cause of obesity.

Endocrine disturbances are a factor in some forms of obesity, but not in all, as is frequently mistakenly supposed. In hyperthyroidism the increased metabolic rate causes loss of weight and the patients are usually thin, but although obesity is observed in some cases of hypothyroidism, all obese subjects do not suffer from thyroid insufficiency. The mechanism of obesity in Cushing's hypophyseal basophilism and in hypercorticalism is not known. Hypophysectomized rats eat little, lose weight, and frequently become cachectic; forcibly fed by a stomach tube, they store more fat than normal controls on the same diet.[1]

Obesity increases the demand on the heart and circulation. The percentage of persons with high blood pressure is higher among the obese than among those with normal weight. Obesity frequently precedes diabetes. Subcutaneous fat hinders the loss of heat by irradiation, and for this reason obese subjects perspire more than thin ones. The average life of obese persons is shorter than that of thin ones and those of normal weight; hence the saying that "as the waistline lengthens the life line shortens."

Treatment of obesity consists mainly in the reduction of the food intake. Exercise may help by increasing the consumption of storage fat, but it should be prescribed in moderation; otherwise it may be harmful. Hundreds of "reducing systems" have been proposed; some are inefficient, and others are dangerous because they restrict the intake of essential foods. A normal adult weighing 130 to 150 lb. who develops moderate physical activity needs 2500 kg.-cal. daily. An obese person must reduce this to 1000 or 1200 kg.-cal. daily to lose weight; occasionally, and only for a short period, the daily caloric intake may be reduced to 600 or 400 kg.-cal. Care should be taken that essential foods such as vitamins, salts, and adequate protein are given. It must be remembered that to lose 1 kg. of fat the equivalent of 7000 cal. must be burned. The most difficult obstacle to overcome in the treatment of obesity is a large appetite. This is easily controlled for a short time, during which weight is lost, but often this control breaks down and fat is again stored.

[1] Samuels, L. T., R. M. Reinecke, and K. L. Bauman, *Endocrinology*, **33**, 87, 1943.

BIBLIOGRAPHY

Castro, J. de, "The Geography of Hunger," Little, Brown, Boston, 1952.

Gillman, J., and T. Gillman, "Perspectives in Human Malnutrition," Grune and Stratton, New York, 1951.

Keys, A., J. Brozek, A. Henschel, O. Mickelsen, and H. L. Taylor, "The Biology of Human Starvation," 2 vols., University of Minnesota Press, Minneapolis, 1950.

Keys, A., and J. Brozek, Body Fat in Adult Man, *Physiol. Rev.*, **33**, 245, 1953.

Mayer, J., Genetic, Traumatic, and Environmental Factors in the Etiology of Obesity, *Physiol. Rev.*, **33**, 472, 1953.

McCance, R. A., and E. M. Widdowson, A Method of Breaking Down the Body Weights of Living Persons into Terms of Extracellular Fluid, Cell Mass and Fat, and Some Applications of It to Physiology and Medicine, *Proc. Roy. Soc., London, s.B.*, **138**, 115, 1951.

McCance, R. A., Overnutrition and Undernutrition. I. Causes, *Lancet*, **265**, 685, 1953.

Medical Research Council, Studies of Undernutrition, Wuppertal, 1946–49, by Members of the Department of Experimental Medicine, Cambridge, Special Report No. 275, London, 1951.

Newburgh, L. H. "Obesity in Clinical Nutrition," Hoeber, New York, 1950.

Osserman, E. P., G. C. Pitts, W. C. Welham, and A. R. Behnke, *J. Applied Physiol.*, **2**, 633, 1950; **4**, 364, 1951 (estimation of water and fat in living persons).

Rynearson, E. H., and C. F. Gastineau, "Obesity," Blackwell, Oxford, 1949.

Widdowson, E. W., R. A. McCance, and C. M. Spray, *Clin. Sc.*, **10**, 113, 1951 (chemical composition of the human body).

CHAPTER 47

The Physiology of Physical Exercise

THE TERM "PHYSICAL EXERCISE" is usually restricted to the mechanical phenomena accompanying the activity of the organs of locomotion, but efficiency in exercise demands the correlated cooperation of all the organs and systems in the body. In studying the physiology of physical exercise it is therefore necessary to consider the organism as a whole made up of integrated parts, the actions of which combine with admirable precision.

The knowledge of the physiology of exercise has many applications. In the first place, work is often a form of exercise. The social importance of work has stimulated the study of this aspect of physiology to such an extent that industrial physiology is now a well-recognized specialty. In the second place, we are living in a mechanical age, and this has created sedentary habits which have made physical exercise necessary for the maintenance of health. To be of benefit, physical exercise must be adapted to the capacity of the individual who practices it. Lastly, an important factor for national defense is the rapid preparation of large forces of men with a high standard of physical efficiency. All the effects of exercise on the different systems and the demands it makes on the organism must therefore be well understood. In this chapter the results of training and the efficiency of the organism as a machine that transforms energy will also be considered, together with other problems of industrial and aviation physiology.

The greater part of the results to be discussed have been obtained from observations and experiments made on man.

OXYGEN CONSUMPTION

Physical examination of a subject who has run a hundred yards shows "shortness of breath," *i.e.*, increased frequency (polypnea), together with some difficulty in breathing (dyspnea). A vigorous rapid heartbeat can be felt by palpating the thorax, and the pulse is more frequent. There are sweating, vasodilatation of the skin, and a rise in body temperature. These are signs of an increased metabolic rate. Muscular contraction is accompanied by an increase in oxygen consumption because, when working, the muscles require a supplement of energy, which is obtained by burning a larger amount of foodstuffs. This increase in combustion causes an increase in the oxygen consumption of the whole organism, and several adjustments in the respiratory and circulatory functions must be made

Table 62. Bodily Changes Produced by Exercise

	Resting	*During exercise*	
		Average	*Maximal*
Oxygen consumption, cc. per min	250	2,500–3,500	5,000
Oxygen debt, liters	....	4–8	16–19
Lactacidemia, mg. per 100 cc. blood	15	50–100	200
Respiration			
Rate	12–16	30	60
Inspiratory volume, cc.	350	2,000	2,200
Ventilation, liters per min.	4.5–6	50–70	120
Circulation			
Pulse rate	70	120–150	200
Systolic output, cc.	60–70	90–110	150
Minute volume, liters	4–5	10–20	35
Systolic arterial blood pressure, mm. Hg	120	160	180
Temperature Δ °C.	....	0.5–1	2

to assure the arrival of more oxygen to the active muscles (Table 62; Fig. 208).

The rate of oxygen consumption in man during exercise increases proportionally to the work performed and can be as much as 10 and 20 times the basal rate. Table 63 summarizes Benedict and Murschhauser's observations in moderate exercise.

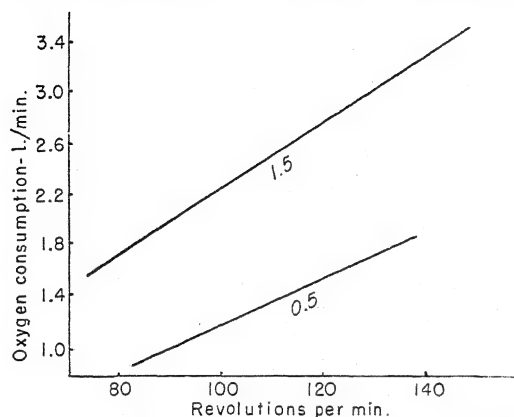



FIG. 208. Oxygen consumption and work performed. In this case work was measured by the revolutions per minute ot a bicycle wheel with loads of 1.5 and 0.5 amp. (*From Benedict and Cathcart, Carnegie Institution, Pub. No. 187, Washington, D.C., 1913.*)

Table 63. Oxygen Consumption and Energy Output in Different Conditions

Condition	*Oxygen consumption, cc. per min.*	*Energy output*	
		Cal. per min.	*Cal. per m.*
Lying down	226–242	1.14	
Sitting	234–260	1.19	
Standing at ease	238–239	1.25	
Standing to attention	....	1.30	
Moving the arms	516	2.53	
Walking			
53 m. per min	667	3.4	0.064
112 m. per min	1296	6.7	0.053
146 m. per min	2240	11.5	0.077

Data taken from BENEDICT, F. G., and H. MURSCHHAUSER, "Energy Transformations during Horizontal Walking," Carnegie Institution, Pub. No. 231, Washington, D.C., 1915.

If the oxygen consumption is measured once a minute during and immediately after moderate exercise and the results are plotted against time, a curve like that of Fig. 209 results. At the beginning of exercise oxygen consumption increases progressively up to a maximum, at which it remains as long as the exercise lasts. When muscular activity comes to an end, the oxygen consumption decreases more or less rapidly, according to the intensity of the exercise performed, and returns to the basal level. The curve can be divided into three parts: (*a*) progressive increase in oxygen consumption; (*b*) steady state; (*c*) recovery.

During the first few minutes of exercise, while the oxygen consumption rises, the oxygen absorbed is less than that needed to oxidize the metabolic products of muscular contraction. This is due to the delay in the adaptation of the respiratory, and especially the circulatory, functions to the greater demand of oxygen by the tissues. Nevertheless the muscles continue to contract, in great part by means of the energy set free by anaerobic chemical reactions. The relative lack of oxygen prevents the complete resynthesis of the lactic acid produced into glycogen. This acid accumulates in the muscle and passes out into the blood; theretore lactacidemia increases. During the steady state a balance is established between the need of oxygen and the speed at which it is delivered to the muscles and used up by them. The amount of lactic acid accumulated neither increases nor diminishes. During recovery the rate of oxygen consumption should fall suddenly to the resting level, but it does not, because the lactic acid which is accumulated during the first stages and which is not removed during the steady state must be disposed of during the period of recovery. The organism has contracted an "oxygen debt," the payment of which is postponed until the exercise has ended. By "oxygen debt" is meant, therefore, the amount of oxygen, in liters, necessary for the removal of metabolic products accumulated while the supply of oxygen is below the needs of the organism. This debt is measured by the amount of oxygen consumed above the basal rate between the end of the exercise and the time when the oxygen consumption has fallen to the resting level.

In violent exercise the oxygen-consumption curve is different. It first increases to a maximum, which is higher than that of the steady state of moderate exercise, but nevertheless does not completely satisfy the oxygen requirements of the organism. The oxygen debt is much

greater and the recovery period more prolonged (Fig. 210).

The exercise an individual can perform is limited by the oxygen supply in one of two ways. In the first place, there is a maximum rate at which oxygen can be absorbed, transported, and used, determined by the efficiency of the respiratory and circulatory adaptations. This maximum is usually 4 liters per min., but in exceptional cases higher figures up to 5.35 liters have been recorded.[1] This maximum rate puts a limit to the speed of exercise; *e.g.*, in a 5,000-yd. race the athlete will have to run at a speed allowing the establishment of a steady state, in which oxygen consumption is sufficient to cover the demand.

In the second place, there is a limit to the oxygen debt that the organism can contract to be paid during the period of recovery. This put a limit to the time during which violent exercise can be performed. In a 100-yd. sprint the runner develops the maximum speed possible, and in the 11 to 12 sec. that the race lasts, only one-sixth of the oxygen necessary can be taken in. The greater part of the muscular effort is made on an oxygen debt, which will be paid during the period of recovery; the larger the maximum possible debt, the longer will the individual be able to sustain the effort. "Oxygen credit" varies from one individual to another. It is large in athletes and can be developed by training up to 16 to 19 liters of O_2 per minute.[1]

FIG. 209. Diagram of oxygen consumption before, during, and after moderate exercise. 1, period of progressive increase; 2, steady state; 3, recovery. Area *A* represents the oxygen needed above the oxygen consumption (oxygen debt). Area *B* represents the oxygen consumption above the actual need (payment of the oxygen debt). *A* and *B* cover the same surface.

FIG. 210. Oxygen need (striped area) and oxygen consumption (curve) in violent exercise.

The intensity of work or exercise can be measured by the rate of oxygen consumption. Dill has suggested a classification into three types of exercise: *moderate*, when the oxygen consumption is up to three times the basal rate; *heavy*, when it is three to eight times the basal rate; and *maximum*, when it is more than eight times the basal rate.

LACTACIDEMIA

When the oxygen demand is greater than the supply, *i.e.*, when there is an oxygen debt, as

[1] ROBINSON, S., H. T. EDWARDS, and D. B. DILL, *Science*, **85**, 409, 1937.

[1] SERGENT, R. M., *Proc. Roy. Soc., London, s.B.*, **100**, 10, 1926.

during the first stages of moderate or in maximum exercise, lactacidemia increases from a resting level of about 15 mg. per cent to 100 and even 200 mg. per cent.

The lactacidemia level represents an equilibrium between the production and removal of lactic acid. One source of lactic acid is its formation in the blood (by glycolysis), which increases in alkalosis and also when the blood passes through the lungs. Another and more important source is muscular contraction, during which glycogen is first broken down to lactic acid and then again resynthesized to glycogen when energy is provided by oxidation (the Pasteur-Mayerhof cycle).

When there is a strong or sustained muscular contraction, the insufficiency of the oxygen supply causes the accumulation of lactic acid, which diffuses into the blood stream. The lactic acid formed by glycolysis or in the course of muscular contraction can be excreted in the urine or sweat, or utilized by the heart, or reconverted into glycogen by the muscle, but the greater part of that which has passed into the blood is taken up by the liver and converted into glycogen (the Cori cycle). Oxygen is needed for this resynthesis.

When the formation of lactic acid does not exceed the capacity of the organism to remove it, as in the steady state of moderate exercise, lactacidemia does not increase. When the effort is greater, during the initial phase while the oxygen supply is below the demand, lactic acid will accumulate in the blood. It will remain at a fairly constant high level if a steady state is established. Later, during the period of recovery, lactacidemia will fall rapidly to the resting level. When the exercise is such that even the maximum possible absorption of oxygen falls short of the requirements, there is no steady state; lactic acid continues to accumulate until the exercise must be discontinued because of fatigue. In this case a maximum oxygen debt has been contracted, and there is a prolonged period of recovery.

There is an evident connection between the lactic-acid level in the blood and the oxygen debt (Fig. 211). At one time Hill believed that an oxygen debt always signified an increase in blood lactic acid and that the excess oxygen consumption of the recovery period was utilized in the removal of the excess lactic acid. It is now admitted that oxygen is used also for the removal of other substances accumulated in the course of muscular contraction. In light exercise there is an oxygen debt without any increase in lactacidemia.[1] It is possible that in this case small amounts of lactic acid or other metabolic products are accumulated in the muscle. Above a certain level of oxygen debt there is a lineal relation between it and lactacidemia (Fig. 211); therefore the capacity of the organism to contract an oxygen debt is related to its ability to limit or tolerate the increase in blood lactic acid.

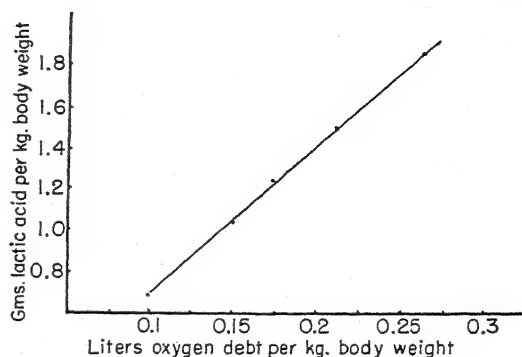



Fig. 211. Lactacidemia and oxygen debt.

The increase in lactacidemia produces several effects:

1. Part of the excess lactic acid is excreted by the kidney. There is therefore an increase in lactic acid and ammonia in the urine collected during and after a period of exercise.
2. Lactic acid displaces CO_2 from the base with which it is combined and forms lactates. This results in an increased respiratory output of CO_2 (see below) and a fall in the alkali reserve.
3. The accumulation of lactic acid in the blood can increase the hydrogen ion concentration and thus stimulate the respiratory center, provoking hyperpnea which results in a speedier elimination of CO_2.

All these adjustments, especially those taking place in the initial period, result in the elimination of a larger amount of CO_2 than that corresponding to the consumption of oxygen, and thus the RQ increases and in some cases reaches values near 2. Respiratory distress (dyspnea) at the beginning of exercise has been attributed to the excess lactic acid. It disappears suddenly

[1] Owles, W. H., *J. Physiol.*, **69**, 214, 1930; Margaria, R., H. T. Edwards, and D. B. Dill, *Am. J. Physiol.*, **106**, 689, 1933.

("second wind") once the necessary adaptation has been accomplished, even though the effort is sustained at the same level. This disappearance of dyspnea is probably due to readjustments in respiratory and circulatory functions and in temperature regulation.

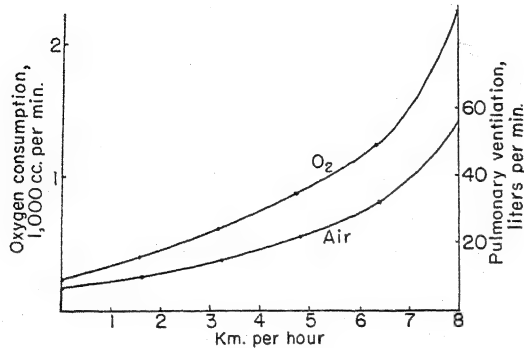



FIG. 212. Pulmonary ventilation and oxygen consumption while walking at different speeds. (*Drawn from Gemmill, "Physiology in Aviation," Charles C Thomas, Springfield, Ill., 1943.*)

RESPIRATORY CHANGES

A visible effect of exercise is the increase in the frequency and depth of breathing. The frequency can rise from 16 to 20 movements per minute in the resting state to 50 or more per minute in violent exercise, and the inspiratory volume can increase from 500 to 2,000 cc. An increased ventilation results, as a result of which the oxygen intake can increase. There is a direct relationship between pulmonary ventilation and work performed, and between pulmonary ventilation and oxygen consumption (Fig. 212).

The mechanism by which this perfect adjustment between oxygen consumption and pulmonary ventilation is carried out is as yet incompletely understood. In the course of physical exercise there are several changes in the chemical composition of the blood which can provoke hyperpnea. Thus, the rise in CO_2 tension, the fall in O_2 tension, and the decrease in pH are factors which stimulate the respiratory center and peripheral chemoreceptors.[1] Breathing air with an excess of CO_2 causes a result similar to that of exercise; first the depth and then the frequency of the respiratory movements increase. Nevertheless, by adding CO_2 to the inspired air it is not possible to provoke such an intense hyperventilation as can be obtained by violent exercise. Dripps and Comroe[1] observed on an average a respiratory volume of 76.3 liters per min. in normal subjects breathing oxygen with 10.4 per cent CO_2. The same subjects raised the respiratory volume to 109.6 liters per min. during maximum exercise. It has also been found that in some types of exercise the respiratory volume continues to increase although the alveolar CO_2 tension falls.

A certain relation between the pH of arterial blood and pulmonary ventilation has been found, although this relation is not a simple one.[2] Gesell[3] maintains that the regulation of the respiratory rate is mainly due to changes in the pH of the respiratory center. In moderate work, however, there is no change in the pH of the blood, and the oxygen tension does not fall.

An increase in the excitability of the respiratory center has been considered by some workers as the principal mechanism in the respiratory adaptation to exercise. The results of certain experiments are in favor of this interpretation; thus, adding 3 per cent CO_2 to the inspired air causes a much greater increase in ventilation during exercise than at rest.[4] The fact that the respiratory rate increases rapidly from the very beginning of exercise, before metabolic products can accumulate in the blood and act on the respiratory center, has been attributed to reflex stimulation of this center. These reflexes have their origin in the active muscles and joints,[5] in the lungs, in the carotid or aortic bodies, and in the great veins and auricles, or are due to substances produced by the muscles acting on chemoreceptors.[6] The excitability of the respiratory center could also be increased by changes in the circulation and therefore in the blood supply of the center,[7] or by the irradiation to the respiratory center[8] of impulses that go from the brain to the active muscles.

[1] HEYMANS, C., J. JACOB, and G. LILJESTRAND, *Acta physiol. Scandinav.*, **14**, 86, 1947.

[1] DRIPPS, R. D., and J. H. COMROE, *Am. J. Physiol.*, **149**, 43, 1947.

[2] ARBORELIUS, M., and G. LILJESTRAND, *Skandinav. Arch. f. Physiol.*, **44**, 215, 1923.

[3] GESELL, R., *Physiol. Rev.*, **5**, 551, 1925.

[4] NIELSEN, M., *Skandinav. Arch. f. Physiol.*, **74**, 299, 1936.

[5] COMROE, J. H., and C. F. SCHMIDT, *Am. J. Physiol.*, **138**, 536, 1943.

[6] ASMUSSEN, E., and M. NIELSEN, *Acta physiol. Scandinav.*, **12**, 171, 1946.

[7] VON EULER, U. S., and G. LILJESTRAND, *Acta physiol. Scandinav.*, **12**, 268, 1946.

[8] KROGH, A., and J. LINDHARD, *J. Physiol.*, **51**, 182, 1917; SCHMIDT, C. F., and J. H. COMROE, *Am. J. Physiol.*, **3**, 151, 1941.

To sum up, nerve impulses arising in the active limbs and the lungs, or irradiated from the cortex to the respiratory center, play an important, but not exclusive, part in the hyperpnea of exercise. The increase in CO_2 and the decrease in pH in the blood, the accumulation of lactic acid, the changes in the circulation occurring in the respiratory center, impulses arising in the carotid and aortic bodies, and even the increase in body temperature, add their effects to those mentioned above and contribute to produce hyperpnea.[1]

CIRCULATORY CHANGES

Muscle performs work by burning fuel. Therefore when work increases, as in exercise, the oxygen supply must also be increased, hence a greater blood flow is needed through the active muscles. Krogh, Rein, Millikan, and others have demonstrated this increase in blood flow, which is prolonged beyond the period of activity. An exception to this occurs when there is a very strong and sustained contraction, in which case the blood flow diminishes or even ceases, but once the contraction is ended the blood flow is considerably increased.

This increase in blood flow is produced mainly by two mechanisms: (*a*) a local vasodilatation; (*b*) a rise in blood pressure. Local vasodilatation of arterioles and capillaries is always accompanied by the opening of a large number of capillaries that were closed in the resting muscle.[2] The number of open capillaries in an active muscle can be up to 10 times those open in a resting muscle. Thus oxygen pressure in the tissues is raised, and oxygen consumption by the muscle increases between 33 and 90 per cent.

There are nervous and humoral mechanisms of vasodilatation. The nervous mechanism consists in a diminished vasoconstrictor tone, produced reflexly by the action of the high blood pressure on the carotid sinus. The humoral mechanisms have local effects of great importance. During exercise the increased CO_2 tension, the decrease in O_2 tension and pH, the local accumulation of metabolic products such as lactic acid, adenylpyrophosphate, etc., cause local vasodilatation in the active muscles and thus increase the local blood flow.

[1] Comroe, J. H., *Physiol. Rev.*, **24**, 319, 1944.
[2] Krogh, A., "The Anatomy and Physiology of the Capillaries," revised ed., Yale University Press, New Haven, 1929.

There is usually an increase in the cardiac output.[1] The minute volume is 4 to 5 liters in the resting subject, but it can rise to 35 liters in violent exercise. This increase is proportional to the work performed per minute, and it is brought about by a faster heart rate and a greater systolic output. When the exercise ceases, the minute volume rapidly falls to the resting level.

Systolic output. The systolic output is 60 to 70 cc. in the resting normal subject, and it can increase to 100 to 150 cc. in maximum exercise performed by trained athletes. The systolic output is dependent on the amount of blood returning to the heart; therefore the venous return must increase during exercise. A complicated mechanism carries this out:

1. There is an increase in the circulating blood volume, produced by the evacuation of stored blood in the skin, splanchnic area, liver, spleen, etc.
2. The venous return is increased by (*a*) the contractions of the skeletal muscles that empty the veins toward the heart, the venous valves preventing its reflux to the periphery on relaxation; (*b*) the more frequent and ample respiratory movements and the contraction of the abdominal muscles that accelerate the circulation in the large veins of the abdomen and the thorax.

This greater venous return increases the strength of the systolic contraction, as the heart muscle fibers are distended by the greater diastolic filling of heart (Starling's law), and the distention of the right auricle reflexly increases the heart rate (Bainbridge's reflex).

Heart rate. The increase in heart rate (Fig. 213) produced by exercise is not due exclusively to the reflex decrease in vagal tone provoked by the distention of the right auricle by the greater venous return (Bainbridge's reflex). The pulse is accelerated as soon as exercise commences; the first cardiac cycle already shows a shortening of the diastole.[2] Probably this is due to a psychic effect on the vagal tone. Bainbridge's reflex and other factors, such as an increase in body temperature, sympathetic nerve impulses, adren-

[1] When the local increase of blood flow produced by local vasodilatation is not considerable, compensatory vasoconstriction of other areas maintains a constant cardiac minute volume.
[2] Bowen, W. P., *Am. J. Physiol.*, **11**, 59, 1904.

aline secretion, etc., act at a later stage. The heart rate increases rapidly and is stabilized at a maximum dependent on the intensity of the exercise, the age of the subject, his physical condition (training), the temperature of the environment, the atmospheric pressure, etc.

Fig. 213. Variations in systolic and diastolic blood pressure and in pulse rate during and after exercise.

The most important of these factors is the intensity of exercise; the heart rate increases with the intensity of the effort made. Exercise of the same intensity causes a greater increase in heart rate in young subjects that in older ones (Table 64). With age the highest heart rate produced

Table 64. Heart Rate in Walking at 5.6 Km. per Hr. on a Gradient of 8.6 Per Cent*

Age, Years	*Heart Rate*
6	170
10	164
14	160
18	150
22	146
26	143
30	140
34	137
38	134
42	134

Source: Dill, D. B., *Am. Heart J.*, **23**, 441, 1942.

* Oxygen consumption increased to seven times the basal rate.

by maximum exercise decreases. From eighteen to twenty-five years this maximum rate is 190 to 210, at fifty it is down to 160 or 170, and at seventy it does not rise above 150.[1] Subjects in training have a smaller increase in heart rate than untrained subjects performing the same exercise. This is due mainly to the greater increase in the systolic output of the subjects in training.

[1] Dill, D. B., *Am Heart J.*, **23**, 441, 1942.

When the temperature of the environment is above 20°C., the heart rate in exercise reaches a higher level (Fig. 214). The barometric pressure influences the heart rate in exercise mainly because of a decrease in the partial pressure of oxygen causing varying degrees of anoxia, the effects of which will be studied later on.

After exercise has ceased, the heart rate diminishes, at first rapidly, then gradually until it falls to the resting level (Fig. 213).

Arterial blood pressure. The systolic blood pressure rises rapidly as soon as exercise is begun and reaches a maximum conditioned by the intensity of the exercise. It then falls a little and remains at a constant level, the height of which is conditioned by the intensity of the exercise. When the exercise is ended, the blood pressure falls rapidly and reaches the resting level before the heart rate becomes normal. The diastolic blood pressure has similar but less marked fluctuations;[1] after exercise it drops rapidly and then again rises. The pulse pressure is greater during exercise, and smaller immediately after, than it is at rest (Fig. 213).

Several factors cause these changes in blood pressure. The peripheral resistance diminishes during exercise[2] owing to vasodilatation in the active muscles, which is not completely com-

[1] Paterson, W. D., *J. Physiol.*, **66**, 323, 1928; Eskildsen, P., H. Gotzsche, and A. T. Hansen, *Acta cardiol.*, **4**, 199, 1949.

[2] Dexter, L. *et al.*, *J. Applied Physiol.*, **3**, 439, 1951.

pensated by vasoconstriction in other areas. The rise in blood pressure is due to an increase in the cardiac minute volume; therefore systolic and pulse pressure increase more than the diastolic pressure. Stabilization at a level slightly below the maximum observed at the beginning of exercise is probably due to compensatory mechanisms, *e.g.*, mechanical stimulation of the pressoreceptors of the aorta and carotid sinus, and stimulation of the chemoreceptors in the aortic and carotid bodies, by variations in the pH and other changes in the blood. Once the exercise is ended, the cardiac minute volume diminishes, and therefore the systolic and diastolic pressures fall. The rise in diastolic pressure a few minutes after the end of exercise has been attributed to constriction of the vascular territories, which were dilated during the exercise.[1]

FIG. 214. Effect of temperature on pulse rate after 10 min. exercise of different intensities. (*Dill, D. B., Am. Heart J., vol. 23, p. 441, 1942.*)

[1] OGDEN, E., and N. SHOCK, *Proc. Soc. Exper. Biol. & Med.*, **33**, 5, 1935.

CHANGES IN THE BLOOD

Several changes are produced in the blood in the course of exercise. The magnitude of these changes is conditioned by the intensity and duration of the exercise. Plasma volume of circulating blood diminishes,[1] without any significant variation in red cell volume.[2] The viscosity, specific gravity, plasma-protein concentration, and red cell count increase. There are two mechanisms mainly responsible for these variations: (*a*) loss of fluid from the blood; (*b*) entrance of stored erythrocytes into the blood stream, a factor that is not of much importance in man. There is some loss of water through the skin (sweating) and the lungs (increased pulmonary ventilation), but the principal cause of the increase in the red cell count is the passage of water from the blood plasma to the tissues. The following factors facilitate this passage: (*a*) accumulation of metabolic products in the active muscles, which causes an increase in the osmotic pressure of the tissue fluids; (*b*) increased capillary permeability, caused by vasodilatation; (*c*) increase in the capillary filtration area, due to the opening of capillaries; (*d*) the rise in blood pressure. The increase in the red cell count, and therefore in hemoglobin concentration, increases the oxygen-carrying capacity of the blood.

FIG. 215. Changes in the concentration of the principal anions of the serum produced by exercise. (*Dill, D. B., J. H. Talbott, and H. T. Edwards, J. Physiol., vol. 69, p. 267, 1930.*)

[1] KALTREIDER, N. L., and G. R. MENEELY, *J. Clin. Investigation*, **19**, 627, 1940.
[2] NYLIN, G., *Am. J. Physiol.*, **149**, 180, 1947.

Variations in the plasma concentration of the principal ions have been studied during exercise in normal subjects (Fig. 215), among whom the famous marathon runner De Mar was the only one in full training.[1] Bicarbonate and lactic acid varied inversely with each other. The degree of fatigue was related to the increase in lactic-acid concentration (Fig. 216).

Fig. 216. Changes produced by exercise in the concentration of lactic acid and bicarbonate (CO_2 combined at 40 mm. $HgCO_2$). Each subject had run 18 to 20 min. on a horizontal treadmill at 9.3 km. per hr. (*Dill, D. B., J. H. Talbott, and H. T. Edwards, J. Physiol., vol. 69, p. 267, 1930.*)

Moderate exercise causes insignificant changes, or none, in the blood sugar; hepatic glycogenolysis compensates for the increased utilization of muscle glycogen. Maximum exercise of short duration causes a 10 to 60 per cent increase in blood sugar,[2] probably due to glycogenolysis induced by sympathetic stimulation and secretion of adrenaline. Maximum and prolonged exercise, *e.g.*, a marathon race, exhausts the hepatic glycogen reserve and thus can provoke hypoglycemia.[3]

[1] Dill, D. B., J. H., Talbott, and H. T. Edwards, *J. Physiol.*, **69**, 267, 1930.

[2] Dill, D. B., H. T. Edwards, and S. Mead, *Am. J. Physiol.*, **111**, 21, 1935.

[3] Levine, S. A., B. Gordon, and C. L. Derick, *J. A. M. A.*, **82**, 1778, 1924.

CHANGES IN RENAL FUNCTION

After violent exercise of short duration (a 400-yd. race) the renal blood flow decreases and usually glomerular filtration and diuresis also decrease.[1] Urea clearance is depressed. The excretion of chloride and sodium diminishes considerably, while that of acids, ammonia, and phosphate increases.[2] Albuminuria is frequently observed, and when the lactic and pyruvic acid concentration in the blood rises sufficiently, these acids are also found in the urine.

CHANGES IN BODY TEMPERATURE

About 75 per cent of the energy used by the organism is converted into heat. This heat is lost through the skin and the lungs. Heat loss is increased in the following ways: (*a*) evaporation of water in the expired air by increased pulmonary ventilation; (*b*) evaporation of water on

[1] White, H. L., and D. Rolf, *Am. J. Physiol.*, **152**, 505, 1948; Chapman, C. B., *et al.*, *J. Clin. Investigation*, **27**, 639, 1948.

[2] Wilson, D. W., W. L. Long, H. C. Thompson, and S. Thurlow, *J. Biol. Chem.*, **45**, 755, 1925.

the skin (sweating); (*c*) increased convection by the movement of the exercising parts.

In exercise, when heat production exceeds the capacity to lose heat, the body temperature rises 1 to 1.5°C. This rise in temperature is not due to insufficient heat loss; it seems to be dependent on the heat-regulating nerve centers.[1] The increase in body temperature depends on the amount of exercise and individual factors such as the efficiency of temperature-regulating mechanisms. In some cases environmental factors, such as temperature and humidity, also play a part. The increase in body temperature facilitates the dissociation of oxyhemoglobin and accelerates metabolic processes, thereby aiding the adaptation of the subject to the conditions of exercise.

LASTING EFFECTS OF EXERCISE

Repeated exercise increases the working capacity of the organism.[2] The object of training is to acquire this increased capacity by increasing (*a*) strength; (*b*) resistance (sustaining an effort); (*c*) accuracy and sureness of movements.[3] The increase in strength is obtained by greater muscular development (hypertrophy and increase in motor units) and perhaps by physical and chemical adaptation: larger stores of glycogen and phosphocreatine, increased ability to dispose of lactic acid, greater activity of the oxidation-reduction mechanisms, etc.

Increase in resistance is conditioned by the capacity to respond to the greater oxygen demand created by exercise. Training results in hypertrophy of the heart and an increase in blood volume.[4] There is a diminished cardiac minute volume at rest, and an increase in the vital capacity of the lung, so that in exercise the minute volume and pulmonary ventilation can increase to a greater extent than in untrained subjects. The alkali reserve also increases, and thus larger amounts of lactic acid can accumulate in the blood. Trained subjects have a lower resting pulse rate than the untrained, because of an increase in vagal tone.

A more specific result of training is sureness and accuracy in the performance of movements. This is due in great part to improved neuromuscular coordination. The muscles respond more rapidly, and there is a more accurate adjustment in the reciprocal contraction and relaxation of synergical and antagonistic muscles. For this reason, to obtain better results in a particular type of exercise, *e.g.*, rowing, it is necessary to practice this type of exercise, besides undergoing the general training. The more accurate coordination of the different movements results in economy of effort, as only the appropriate muscles are contracted and only in the necessary measure; thus fatigue is delayed. The recovery period is also much shorter in trained subjects than in untrained ones.

Overtraining causes an abnormal condition in which several symptoms (lack of interest and initiative, sleepiness, irritability, etc.) and signs (tachycardia, arrhythmia, loss of weight, etc.) are prominent. This condition is usually due to a too-rapid increase of the effort demanded. The sudden interruption of training can also be harmful.

Measurement of physical fitness. The capacity to perform work is difficult to measure, because it is conditioned by many physical, psychic, and environmental factors. The appraisal of physical fitness is nevertheless of great importance in the physiology and hygiene of work, sport, and military training. This appraisal can be made by comparing the effect on certain physiological processes of (*a*) moderate exercise, and (*b*) maximum exercise, in subjects in good and in bad physical condition.

The performance of moderate exercise which can be kept up for a long time provokes in physically fit subjects less oxygen consumption, less marked rises in blood pressure and heart rate, lower increase in blood lactic acid and a larger systolic output, than in subjects in poor physical condition. After the exercise has ended, blood pressure and heart rate return to basal levels in less time in subjects in good physical condition than in those whose condition is poor.

Maximum exercise is kept up for a longer time by a subject physically fit than by one in poor condition. The former has a greater oxygen consumption, lower heart rate, greater systolic output, higher lactacidemia, and returns more rapidly after the exercise has ended to a basal blood-pressure and heart-rate level.

Subjects in good physical condition can therefore perform moderate exercise with small

[1] Pellegrini, A., G. Riva, and R. Margaria, *Boll. Soc. ital. biol. sper.*, **22**, 474, 1946; Asmussen, E., and M. Nielsen, *Acta physiol. Scandaniv.*, **14**, 382, 1947.

[2] Steinhaus, A. H., *Physiol. Rev.*, **13**, 103, 1933.

[3] Lindhard, J., *Ergebn. d. Physiol.*, **33**, 337, 1931.

[4] Kjelberg, S. R., V. Rudhe, and T. Sjöstrand, *Acta physiol. Scandinav.*, **19**, 152, 1949.

adjustments in their physiologic equilibriums, and they can maintain a steady state at a higher level during a longer time than is possible for subjects in poor condition. Recovery of basal conditions is also more rapid in subjects in good condition than in those in bad condition.

Several methods of measuring physical fitness have been proposed.[1] The duration or the amount of work performed in maximal exercise can be measured, or the changes provoked by a certain amount of work on the oxygen consumption, lactacidemia, oxygen debt, blood pressure, etc., can be determined. Usually the heart rate is counted during and after exercise.

THE ENERGETICS OF EXERCISE

Muscles are machines that convert chemical energy into mechanical work and heat. An outstanding feature of exercise is the great amount of energy the organism can develop. For example, a subject weighing 75 kg. in complete rest has an output of about 1.2 cal. per min.; this output is increased to 10 or 15 cal. per min. in violent exercise.

The efficiency of the organism considered as a machine for converting energy can be measured by determining the energy output (either in a calorimeter or by calculating it from the oxygen consumption) and the work performed. The net maximum efficiency in man is around 25 per cent (Table 65).

Table 65. Maximum Efficiency of Engines, Horse, and Man

	Efficiency, %
Steam engine	15
Gasoline motor	18
Horse	24
Man (climbing stairs)	25
Diesel motor	35

Source: BRODY, S., and E. A. TROWBRIDGE, Bull. 383, Agricultural College, University of Missouri, 1937.

To obtain the net efficiency, the energy consumption at rest (BMR, basal metabolic rate) must be subtracted from the total energy consumption. Thus gross efficiency is

$$E_g = \frac{100W}{H}$$

where W is work performed and H is total energy. Net efficiency is

$$E_n = \frac{100W}{H - BMR}$$

[1] TAYLOR, C. L., *Ann. Rev. Physiol.*, **7**, 599, 1945.

Net efficiency is conditioned by the type of work done (ergometer, treadmill, carrying a load, running, etc.), training, coordination of movements, speed, etc. For example, Fig. 217 illustrates the effect of speed on efficiency. The curve responds fairly accurately to Hill's equa-

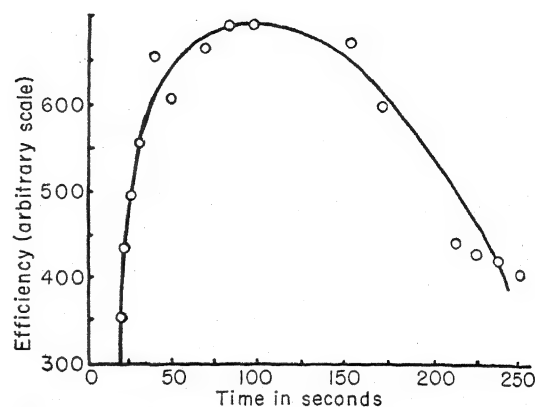



FIG. 217. Efficiency and speed of exercise. The subject climbed a staircase; the greatest efficiency corresponded to a speed of 50 steps per minute. (*Lipton, H., J. Physiol., vol. 57, p. 337, 1923.*)

tion for efficiency and duration of muscular contraction, and it shows that maximum efficiency is obtained only within a fairly narrow range of speed.[1] These results have important applications in the physiology of exercise; *e.g.*, walking is most economical when performed at a rate 120 paces per minute.

Sources of muscular energy. At one time protein was thought to be a source of energy for muscular contraction, because the uric acid and creatinine excreted in the urine increase during heavy exercise. Proof has now been obtained that these nitrogen compounds are simply a sign of muscular wear and tear caused by contraction. In any case, protein could supply only a very small part of the energy used in performing mechanical work. Carbohydrate and fat must therefore be considered as possible sources of energy.

Observations on the respiratory quotient have given valuable information in this respect. The respiratory quotient increases in exercise, in relation to the intensity of exercise (in violent exercise the RQ is above 1 because of the excess CO_2 eliminated). This shows that there is a greater consumption of carbohydrate. Never-

[1] HILL, A. V., *Lancet*, **281**, 947, 1951.

theless, in prolonged exercise the RQ seldom rises to the value of 1, and when the diet is poor in carbohydrate, there not only is a low RQ when resting, but also the rise of the RQ provoked by exercise is less (Table 66).

Table 66. The Effect of Exercise on the Respiratory Quotient

Previous diet	*At rest before exercise*	*During exercise*	*After exercise*
Carbohydrate..........	0.85	0.90	0.78
Carbohydrate-poor......	0.79	0.82	0.75

Source: BENEDICT, F. G., and E. P. CATHCART, Carnegie Institution, Pub. No. 187, Washington, D. C., 1913.

A carbohydrate-rich diet is therefore advisable for increasing the capacity for work. Moreover the maximum RQ diminishes in successive periods of exercise. All these results have led to the conclusion that not only carbohydrate but also fats are sources of energy in muscular contraction. The way in which muscle obtains energy from fat is not yet known; it is still doubtful whether muscles can utilize storage fats directly, or whether these must first be modified or converted into carbohydrate.[1]

Appetite for fat and protein is greater in athletes and persons doing heavy manual work than in those with sedentary habits. An increased desire for fats is perhaps due to the fact that these substances are an economical source of energy. The larger consumption of protein is perhaps caused by the high vitamin B complex content of meat. Individuals in training are easily fatigued and their capacity for work is small if they are fed a diet poor in vitamin B complex. Normal capacity for work is restored by supplementing the diet with B complex preparations.[2]

Alcohol, caffein, cocaine, strychnine, and nitroglycerine have been tested in trained athletes, but they do not raise the efficiency of persons in good condition.[3]

[1] COURTISE, F. C., and C. G. DOUGLAS, *Proc. Roy. Soc., London, s.B.*, **119**, 381, 1936; DOUGLAS, C. G., and A. C. E. KOCH, *J. Physiol.*, **114**, 208, 1951.

[2] BARBORKA, C. J., E. E. FOLTZAND, and A. C. IVY, *J. A. M. A.*, **122**, 717, 1943.

[3] ASMUSSEN, E., and O. BOJE, *Acta physiol. Scandinav.*, **15**, 109, 1948.

FATIGUE

Fatigue results from muscular activity and from the activity of the nervous system (mental work, etc.). Frequently both types of fatigue combine in varying proportions. The consequence is a diminished capacity for work.

In muscular exercise fatigue is due to the accumulation of metabolic products, the exhaustion of energy-yielding substances, and other muscular, respiratory, and circulatory phenomena. Mental fatigue is more difficult to appreciate objectively. The following can be mentioned among causes of this type of fatigue: anxiety, nervous tension, monotony of the task, noise, vibrations, lack of adequate rest periods, etc.

INDUSTRIAL PHYSIOLOGY

The results obtained by observing the physiologic changes in the organism produced by exercise have been applied to the study of working conditions in factories, the army, etc. Thus a new field has been opened, that of industrial physiology. Its main object is to determine the optimal conditions for obtaining maximum efficiency and for retarding the onset of fatigue. This is of great importance in industry. The great variety of types of work in industry and the environmental circumstances (climate, light, ventilation, etc.) in which it must be performed present many problems to the physiologist.

Each type of work has its particular problems, and several attempts have been made to classify work. In one classification three types of work are differentiated according to whether strength (stone breaking, foundry), endurance (common labor), or speed (packing, typewriting) is the predominant feature. The environmental factors (barometric pressure, temperature, humidity, etc.) are quite as important as the type of work. Many examples could be given of the applications of physiological knowledge to the solution of problems of this nature. For example, laborers working on the Boulder Dam in hot weather frequently suffered from cramps. A physiological study of these workers showed that they sweated profusely and therefore lost water and sodium chloride in great amounts. They drank large quantities of water, thus replacing it but not the salt. The administration of salt immediately cured the cramps and prevented not only considerable suffering but also

the loss of many working hours with the consequent economic disturbance.

This branch of physiology received great impetus during the Second World War, not only with respect to work in the factories, which was of fundamental importance, but also with respect to the activity of soldiers, because the variety of battle fronts in all latitudes created many problems. The fighting conditions of an infantry soldier in the Russian winter were widely different from those of a tank driver in the African desert or of a marine disembarking on an island in the Pacific. The advance of aviation has been a special source of problems for physiology and medicine.

One of the aims of industrial physiology is to find out what kind of work is suited to the capacity of the workman, his physical condition, and the environment in which he works. If the effort is an excessive one, it will be harmful; if it is below the workman's capacity, it will be economically inefficient. In this respect training and fatigue have a different significance in industry and in exercise. Training can be directed to increase strength and endurance, but it will give better results if its object is to acquire perfection in the coordination of movement. Ability depends to a great extent on inborn factors, but its development by training has a considerable stability. As Abramson says, "Man needs to learn how to ride a bicycle only once in his life, while strength and endurance must be built up each year."

Psychological factors are quite as important as physical ones in the onset of fatigue, if not more so. Not only the conditions in the factory (duration and nature of the work performed, speed, rhythm, etc.), but also those outside the factory (rest, home conditions, etc.) are important. Fatigue can be delayed by many devices, such as adequate rest periods, music, competition between groups, knowledge of the significance of the task, etc.

AVIATION PHYSIOLOGY[1]

The remarkable technical progress in aviation has created many problems in physiology. Airplanes are built that can climb more than 40,000 ft. (12,000 m.) in a few minutes, fly at 600 miles (1,000 km.) an hour, and nosedive at speeds near those of sound. Jet-propulsion planes that can fly above the atmosphere may be built in the near future. Already planes have flown at heights above 70,000 ft. (22,000 m.), *i.e.*, an altitude equivalent to 95 per cent of the atmosphere. This type of flying has presented new problems, *e.g.*, rapid displacement through the atmosphere when ascending or descending (heating of the plane, acceleration, deceleration). Moreover, certain vital properties of the atmosphere no longer exist: there are no atmospheric gases; ultraviolet and cosmic rays are not filtrated; the plane is in a dark vacuum and may collide with meteorites. The pilots of these planes must be especially protected to be able to resist the conditions of flight. Anoxia, sudden changes in barometric pressure, cold, and acceleration are some of the factors creating outstanding problems that will be considered here.

Anoxia. Historically anoxia was the first obstacle man had to overcome in his conquest of the air. Remarkable experiments by Paul Bert with a pneumatic pump, performed on animals and man, proved that the disorders observed in the course of balloon ascensions were due not to the fall in barometric pressure, but to the fall in the partial pressure of oxygen—therefore, to lack of oxygen.

The conditions governing the combination of hemoglobin with oxygen at different partial pressures of oxygen (see Sec. III, Respiration) are such that up to a height of 13,000 ft. (4,000 m.) there are no serious effects on the organism. At 20,000 ft. (6,200 m.) the atmospheric pressure is 380 mm. Hg, one-half that at sea level, and the oxygen partial pressure is also half the normal (Fig. 138). At this oxygen pressure the blood is insufficiently oxygenated and signs of anoxia are observed. To prevent anoxia the oxygen pressure must be increased. This can be done in one of two ways: (*a*) by increasing the percentage of oxygen in the air breathed by the pilot from an oxygen mask, such as the widely used Boothby, Lovelace, and Bulbulian (BLB) model or the more elaborate ones with regulatory valves; (*b*) enclosing the pilot in a hermetically sealed cabin or in a special suit in which air, more or less rich in oxygen, is kept at high pressure. As a safety measure the military authorities in the Second World War

[1] Armstrong, H. C., "Principles and Practice of Aviation Medicine," Williams & Wilkins, Baltimore, 1939; Gemmill, C. L., "Physiology in Aviation," Charles C Thomas, Springfield, Ill., 1943.

obliged flying personnel to inhale oxygen at any height above 13,000 ft. (4,000 m.).

Gaseous embolism. This is another physiological problem created by technical progress. The rapid ascent of modern planes is accompanied by a correspondingly rapid fall in barometric pressure. This causes nitrogen dissolved in the blood to be released in gaseous form, and the bubbles thus produced can clog the small blood vessels (gaseous embolism).

The blood and the body fluids are always saturated with the gases of the atmosphere, because the blood is exposed to these gases as it passes through the lungs. They are dissolved in the plasma according to their solubility and partial pressure. As nitrogen forms 79 per cent of the air, a relatively large amount of it is dissolved in the blood (1.7 cc. of N_2 per 100 cc. at 1 atmosphere). An equilibrium is established between the N_2 in the air, the blood, and the body fluids. When the barometric pressure increases (in divers it can be up to 5 and even 7 atmospheres) the partial pressure of N_2 increases in the inspired air and in the lungs, and therefore also in the blood and in the tissues, and a new equilibrium is established. If the barometric pressure then falls gradually to the normal level, the nitrogen pressure will be higher in blood than in the air, and N_2 leaves the blood as it passes through the tissues. If the barometric pressure falls gradually, no disturbance occurs, but if it falls suddenly so that the partial pressure of N_2 in the tissues is twice that in the air, N_2 is released in gaseous form and bubbles appear in the blood. As N_2 is more soluble in fats than in water, bubbles are more easily formed in tissues with a large fat content, *e.g.*, the central nervous system. These bubbles are a serious obstacle to the passage of blood in the small arteries, and the areas nourished by the obstructed vessels are the site of disturbances of lesser or greater severity. At first there is a sensation of itching in the skin, then pains of increasing severity in the chest and joints, shooting pains along the nerves, dyspnea, cyanosis, coughing, paralysis, etc. These signs and symptoms have often been observed in divers coming up too rapidly to the surface after having been submerged for a certain time at a pressure of 4 to 5 atmospheres (caisson disease). It is therefore necessary to raise them to the surface gradually so they can eliminate the excess N_2 dissolved and a new nitrogen equilibrium can be established between the air and the tissues without the formation of N_2 bubbles. This method was proposed by Haldane many years ago and gives excellent results, but the equilibrium is reached only after a relatively long time. More recently the breathing of air in which N_2 has been replaced by oxygen or helium has been proposed.

An aviator ascending rapidly is in the same situation as a diver coming up to the surface. For example, at 42,000 ft. (12,800 m.) the barometric pressure is 150 mm. Hg or about one-fifth of 1 atmosphere. Gases dissolved in the blood at the normal pressure will be eliminated, and a new equilibrium will be established at this barometric pressure, as was stated above. Oxygen and CO_2 rapidly combine with hemoglobin and the alkaline reserve, but N_2 must go from the tissue fluids to the plasma and from thence to the alveolar air. If there is a very rapid ascent, as frequently occurs in fighting conditions, the equilibrium between N_2 in the tissues and the air cannot be established gradually, so this gas is released and forms bubbles, with the results described. Cabins that can be kept at constant pressure have solved some of these problems, but they may accidentally lose pressure.

Zuntz suggested many years ago that by breathing pure oxygen part of the N_2 dissolved in the body fluids could be rapidly eliminated. In 1 hr. about 50 per cent of the dissolved N_2 is displaced, but it takes over 5 hr. to displace the remaining 50 per cent. This is due to the fact that dissolved nitrogen is much more rapidly displaced from the body fluids than from the tissues, especially those with a high fat content, where there is slow circulation and a large amount of gas dissolved (adipose tissue). Exercise while breathing oxygen hastens the elimination of N_2 because it speeds up the circulation (increase in cardiac minute volume, decrease in circulation time, etc.). A pilot can thus eliminate a large proportion of the nitrogen dissolved in his body fluids before going up and can thus diminish the risk of gaseous embolism. There are individual differences in this respect, which can be demonstrated in low-pressure chambers.

Disturbances in the ear. In the course of one of the first ascensions in a balloon, December 17, 1873, the pilot felt acute pain in the ears. Anyone who has traveled in an airplane, or even gone up or down in a rapidly moving elevator, has experienced such disturbances.

The pain is due to the difference between the pressure on the outside of the eardrum and the pressure in the middle ear, *i.e.*, on the inside of the eardrum. The eustachian tube, which goes from the middle ear to the pharynx, is usually closed but opens on swallowing or yawning. When an airplane rises, the pressure on the outside of the eardrum falls below that in the middle ear, and the tympanic membrane bulges out; this distention is the cause of the pain. When the difference in pressure between the middle ear and the external environment is about 15 mm. Hg, the eustachian tube is opened, and thus the pressure on each side of the eardrum becomes the same. A click is heard when this occurs. It is enough to perform the simple act of swallowing to prevent all this. In the opposite case, *i.e.*, when the atmospheric pressure is greater than that of the middle ear, as when descending in an airplane or on diving into water, the eustachian tube acts as a valve and does not allow the passage of air from the mouth to the ear; therefore if the eustachian tube is not opened by swallowing, yawning, or crying out, the tympanic membrane will be more and more pushed in toward the internal wall of the middle ear. If the descent is rapid enough (about 3,000 ft. per min.) the tympanic membrane can be ruptured. A noise like that of an explosion is heard, and a sensation of having been knocked on the head is experienced.

Airplanes when nosediving descend at a tremendous velocity. If the pilot does not keep his eustachian tube open by repeated swallowing, or if he has a catarrh that obstructs it, the tympanic membrane will be ruptured, pain will be felt, and a more or less severe hemorrhage may occur, with a resultant deafness. For this reason pilots should not fly when suffering from a cold, and if they are obliged to do so, they should ascend or descend gradually to avoid sharp differences in pressure between the atmosphere and the middle ear.

Cold. Another problem that must be considered is the protection of the aviator against cold. For every 1,100 ft. (330 m.) of ascent, a decrease in temperature of about 2°C. takes place, up to 38,000 ft. (11,500 m.), the upper limit of the troposphere; at higher altitudes (stratosphere) the temperature remains constant at −55°C. The organism reacts against cold by intense cutaneous vasoconstriction; by diminishing the secretion of sweat, thus decreasing heat loss; and by involuntary muscular contractions (shivering), thus increasing heat production. When there is intense cold or prolonged exposure, symptoms of depression, indifference, and lethargy appear, and finally the subject dies. As fighting planes must rise rapidly to high altitudes where their efficiency is greater, it has been necessary to solve the problem of keeping the pilot warm while flying or parachuting at a height of 32,000 to 42,000 ft. (10,000 to over 12,000 m.) where the temperature is between −45 and −55°C. Especially heated suits have been used for this purpose.

Effects of acceleration. The tremendous velocities developed by jet-propelled planes have created very serious problems. Planes are now built with a speed exceeding that of sound (800 miles per hour); but change in speed, *i.e.*, acceleration, not speed itself, is the cause of the most serious problems in aviation physiology. The results of acceleration can be instantaneous, serious, and even dramatic.

The greatest degrees of acceleration occur when there is a change in the direction of flight, and they are due to pressure on the wings of the machine. For example, when a plane nosedives the velocity is more or less uniform while the plane flies in a straight line, but when the direction of flight changes there is an increase in weight, due to the new force applied at right angles to the original direction. The same thing occurs in "inward" looping. The resultant force of this acceleration is transmitted parallel to the longest axis of the pilot's body when he is in the usual position in a plane—in this case, from the seat to the head. This causes, among other disturbances, an accumulation of blood in the abdomen and a decrease of the flow of blood to the heart and head. The result is a fall in blood pressure and cerebral anemia. The visual apparatus, retina, optic nerves, and nerve centers are particularly sensitive to lack of blood, so an outstanding result of acceleration is a visual defect, consisting in loss of vision known as "blackout" (called *Vorhang*, curtain, by the Germans). If the cerebral anemia is sufficiently prolonged, the pilot loses consciousness. In "outward" looping, on the contrary, blood accumulates in the head and the pilot "sees red" ("red-out"). The acceleration most frequently endured by military pilots is equal to four times the force of gravity and is expressed as 4*g*. An ordinary pilot can endure 3 to 6*g*, while those

with considerable experience can tolerate up to 7 and 8*g*.

When the force of acceleration is transmitted at right angles to the main axis of the body, its effects are not so serious. Thus if the pilot lies down during the descent, when the plane straightens out the force of acceleration will act in a transverse direction to the long axis of the pilot's body. Hence the adoption of this position would solve in great part the difficulties encountered; unfortunately it is not easy to control the machine when lying down. The blackout is due mainly to the accumulation of blood in the abdominal viscera. Pilots are therefore advised to bend forward when nosediving, pressing the chin against the chest, and to cry out. These actions if forcibly carried out are accompanied by widespread muscular contraction, including the abdominal muscles, and congestion in the head. Abdominal pressure can also be increased by compressing the abdomen with a special pneumatic belt, or by preventing venous stasis in the lower limbs, but the devices so far used for these purposes are not very satisfactory.

The physiological problems resulting from the constant progress of aviation have brought about a race between the physiologists and the engineers. The new machines submit the pilots to increasingly severe abnormal conditions: great speed, rapid ascent, nosedives, centrifugal force, acceleration that amounts to several times the force of gravity, prolonged flight, very high altitudes with the consequent low temperature and low oxygen pressure, and finally rapid changes in barometric pressure. Physiologists have studied the effects of these abnormal conditions on the organism and have found ways of preventing most of the accidents that these conditions can provoke.

BIBLIOGRAPHY

Atzler, E., Arbeitsphysiologie, *Ergebn. d. Physiol.*, **40**, 325, 1938.

Bainbridge, F. A., A. V. Bock, and D. B. Dill, "The Physiology of Muscular Exercise," Longmans, New York, 1931.

Catchpole, H. R., and I. Gersh, Pathogenic Factors and Pathological Consequences of Decompression Sickness, *Physiol. Rev.*, **27**, 360, 1947.

Comroe, J. H., The Hyperpnea of Muscular Exercise, *Physiol. Rev.*, **24**, 319, 1944.

Dill, D. B., The Economy of Muscular Exercise, *Physiol. Rev.*, **16**, 263, 1936.

Dill, D. B., "Life, Heat and Altitude," Harvard University Press, Cambridge, Mass., 1938.

Eggleton, M. G., "Muscular Exercise," Routledge, London, 1936.

Gemill, C. L., The Fuel of Muscular Energy, *Physiol. Rev.*, **22**, 32, 1942.

Gould, A. G., and J. A. Dye, "Exercise and Its Physiology," A. S. Barnes, New York, 1935.

Ivy, A. C., High Altitude Problems in Aviation, *Federation Proc.*, **5**, 319, 1946.

National Research Council (Committee on Work in Industry), "Fatigue of Workers; Its Relation to Industrial Production," Reinhold, New York, 1941.

Schneider, E. C., and P. V. Karpovich, "Physiology of Muscular Activity," 3d ed., Saunders, Philadelphia, 1948.

Steinhaus, A. H., Chronic Effects of Exercise, *Physiol. Rev.*, **13**, 103, 1933.

Vernon, H. M., Industrial Fatigue in Relation to Atmospheric Conditions, *Physiol. Rev.*, **8**, 130, 1928.

CHAPTER 48

The Body Temperature and Its Regulation

The significance of animal heat. The evolution of heat in animals can be considered from two points of view: (*a*) heat is the final and irreversible form in which energy obtained from the oxidation of foodstuffs is eliminated from the body; (*b*) a certain temperature level is a necessary condition for the maintenance of an adequate rate in biologic processes.

In the resting organism all the energy obtained from oxidations is finally converted into heat. Thus, if the equivalent of 1800 kg.-cal. of foodstuffs is oxidized, 1800 kg.-cal. will be eliminated as heat, when the body temperature is neither rising nor falling. The performance of mechanical work requires additional chemical energy. Only 20 to 25 kg.-cal. out of every 100 of this additional energy is converted into work; 75 to 80 kg.-cal. is lost as heat.[1] For example, the heat production of a subject at rest is 2000 kg.-cal. in 24 hr. During the same period he performs work and for this purpose 1000 kg.-cal. is spent, bringing the total up to 3000. If the mechanical efficiency is 25 per cent, 750 of the additional calories will be given out as heat and only 250 will be converted into mechanical energy.

The maintenance of a certain temperature is of vital importance to the organism, because all biologic processes are conditioned by temperature. At 0°C. these processes slow down considerably, or even stop; as the temperature rises, their velocity increases up to a maximum at an optimum temperature. Higher temperatures provoke disturbances and even death. In mammals the central nervous system ceases to function at 44 to 45°C. and the heart stops beating at 48°C. The van't Hoff–Arrhenius law, which governs chemical reactions,[1] is valid for physiologic processes; *e.g.*, a rise in temperature of 10°C. causes a twofold to threefold increase in pulse rate, oxygen consumption, etc. The law can be applied to animals only within certain limits, because high temperatures damage the tissues (Figs. 218 and 219).

[1] The efficiency of an organism considered as a machine is given by the ratio of the energy converted into work (W) to the total energy evolved (H) minus resting metabolism (BMR). Net mechanical efficiency, $E_n = \frac{100W}{H - BMR}$, varies from 20 to 25.

[1] A rise of 10° C. increases the velocity of chemical reactions two to three times, *i.e.*, the temperature coefficient (Q_{10}) is from 2 to 3.

Two magnitudes must be considered when dealing with heat: (*a*) quantity of heat, measured in calories, which can be summated; (*b*) temperature, measured in degrees of a centigrade or Fahrenheit scale, which cannot be summated. For example, if 1 liter of water at 37°C., holding therefore 37 kg.-cal., is added to another 1 liter of water at 37°C., there will be 2 liters of water holding 74 kg.-cal., at 37°C.—not at 74°C The specific heat is the amount of heat absorbed by 1 gram of a body when its temperature is raised 1°C. The specific heat of water is considered as 1; that of the human body is 0.83. Variations in heat intensity, *i.e.*, temperature, and not in heat quantity, modify the rate of physiologic processes.

According to their capacity to regulate their body temperature, animals are classified *poikilothermic* and *homoiothermic*. The body temperature of poikilothermic animals varies with the temperature of the environment; thus their vital processes are easily modified by changes in the environmental temperature. Inverte-

brates, fishes, amphibians, and reptiles are poikilothermic; their temperature is usually a few degrees above that of the environment. Homoiothermic animals maintain their temperature constant within narrow limits, and changes in environmental temperature provoke only slight oscillations in their body temperature. This constancy is achieved by the regulation of heat production and heat loss. Homoiothermic animals are similar to a thermostat, the temperature of which is kept constant by regulating the mechanism that heats it and the loss of heat from it. Newborn homoiothermic animals are less resistant to cold than adults, because the nervous system is not completely developed at birth. For example, white rats a few days old placed at 3 to 9°C. cease to consume oxygen, their hearts stop beating, and they remain without movement. If they have not been kept at this low temperature for more than a few hours, they will return to their normal condition a few minutes after being warmed. Hibernating animals are homoiothermic from spring to fall, but in winter when the environmental temperature is below 10°C. (50°F.) their body temperature falls to a level little above the outside temperature, their vital processes slow down, and they are lethargic until the following spring. They are awakened from their winter sleep when the outside temperature rises above 15°C. (59°F.). They also wake when it falls below 2°C. (35.6°F.).

FIG. 218. Effect of changes of temperature on the oxygen consumption of smooth muscle (isolated guinea pig uterus). The dotted line is the theoretical curve for an increase in O_2 consumption of 2.5 times for every temperature increase of 10°C. (*Lovatt-Evans, C., "Starling's Principles of Human Physiology," 7th ed., J. & A. Churchill, London, 1938.*)

THE TEMPERATURE OF THE BODY

In medical practice body temperature is measured by placing a thermometer under the tongue, in the axilla, or in the rectum. Rectal temperature[1] in man has a daily oscillation from 35.9 to 37.3°C. (96.5 to 99.1°F.); the average is usually given as 37.0°C. (98.6°F.), and even 38°C. (100.4°F.) is not abnormal in warm climates. Rectal temperature is more reliable

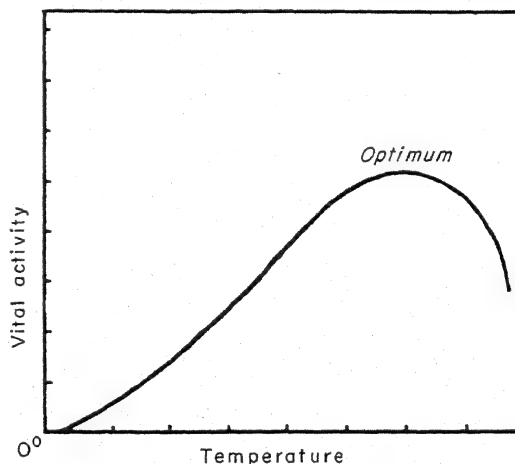



FIG. 219. Curve representing activity of vital processes. There is an increase up to an optimum temperature, above which activity diminishes; at even higher temperatures the cells are damaged and die.

[1] The thermometer must be introduced 6 cm. or more into the rectum to obtain reliable data. Rectal temperature of birds is 40 to 42°C. (104 to 107.6°F.); it is lower in mammals, 35 to 40°C. (95 to 104°F.), usually 39°C. (102.2°F.); in the dog and cat it is 38.5 to 39°C. (101.3 to 102.2°F.); in the horse, 37.5 to 38°C. (99.5 to 100.4°F.); in the ox, 38.2 to 38.7°C. (100.7 to 101.6°F.)

than axillary or buccal temperature, but for esthetic reasons it is not usually taken. Nevertheless when there is any doubt of the existence of fever the rectal temperature should be taken; *e.g.*, in certain cases of tuberculosis it is possible to discover a febrile condition only by taking the rectal temperature every two hours.

If the temperature is taken in the axilla, the thermometer should be left in place for 5 or even 10 min. with the arm held tightly against the thorax. In small children it is easier to take the temperature in the groin. The temperature of the axilla is usually 0.2 to 0.4°C. (0.36 to 0.72°F.) below the temperature in the mouth and 0.5 to 1°C. (0.9 to 1.8°F.) below that of the rectum. It varies daily between 36.2 and 36.9°C. (97.2 and 98.4°F.), with an average of 36.6°C. (97.8°F.), exceptionally reaching to 37.2°C. (98.96°F.). The temperature in the mouth is taken by placing the thermometer under the tongue and keeping the mouth closed; it is 0.3 to 0.5°C. (0.54 to 0.90°F.) below rectal temperature. If hot or cold beverages have been taken, the temperature of the mouth may be above or below normal for a short time.

The temperature of the skin varies in different parts of the body. The temperature of the viscera is 1 to 1.5°C. (1.8 to 2.7°F.) above the temperature in the axilla; it is highest in the liver. The temperature of arterial blood in the limbs has been measured by introducing into the artery a thermocouple in a plastic tube. It varies in different arteries and is modified by the temperature of the environment, the pulse frequency, and the position of the limb.[1] The temperature of venous blood varies more; in the veins of the skin and limbs it is lower than in the arteries. In the vena cava it increases as the blood from the viscera enters the vein and reaches a maximum at the level of the suprahepatic veins. The blood is cooled as it passes through the lung; for this reason it is warmer in the right ventricle than in the left ventricle.

The temperature of the skin varies with the temperature of the environment, the humidity of the atmosphere, the ventilation, and the amount of clothing, on one hand, and with the state of the circulation in the skin and the evaporation of sweat, on the other.

Diurnal variation. The temperature of man has a typical daily variation. It is at its maximum between 5:00 and 8:00 P.M. and at its minimum between 2:00 and 6:00 A.M. Subjects with fever usually have a maximum temperature in the evening also, but in some cases this is not so; the peak will then be discovered by taking the temperature every two hours. The temperature curve is related to variations in muscular activity and metabolism, which are at a maximum in the evening and at a minimum in the early morning. An inversion of the usual cycle is seen in some cases, *e.g.*, in some nightworkers or in subjects traveling around the world, who continue with their usual rhythm, although the day in the place from which they started corresponds to the night in the antipodes. Night birds have a maximum temperature at night.

Other temperature variations. Body temperature is not very stable in the newborn; a hot or a cold bath causes a rise or a fall in temperature. Stability increases with age and is at its maximum in normal adults. In old age the temperature tends to be lower. In women there is a variation related to the menstrual cycle. At the moment of ovulation the temperature falls (about 0.2°C.). It rises 0.1 to 0.4°C. during the second half of the cycle, owing to the effect of the luteal hormone. During menstruation it again falls (0.6°C.). There is a subnormal temperature in subjects with a low metabolic rate, *e.g.*, in undernourishment, starvation, myxedema, or hypophyseal insufficiency. The body temperature is frequently a little above normal in hyperthyroid subjects, owing to high muscular tone and tremor. It is also high on very hot days. It rises a little after a meal and during exercise. Strenuous exercise can produce a transitory rise up to 39 or 40°C. In some cases of tetanus, skin temperature continues to rise for some time after death.

The lower limit of body temperature compatible with life in man is 27 to 29°C. (80.6 to 84.2°F.). Exceptional cases have been reported of survival after the body temperature had fallen to 24 and even 20°C. (75.2 and 68°F.). An increase in temperature to 42 or 43°C. (107.6 or 109.4°F.) is seldom observed. A rise of temperature to 44 or 45°C. (111.2 or 113° F.) produces fatal disturbances in the central nervous system.

THE REGULATION OF THE BODY TEMPERATURE

The body temperature results from the balance of heat produced (thermogenesis) and heat

[1] BAZETT, H. C., *Am. J. M. Sc.*, **218**, 483, 1949; EICHMA, J. W., *et al.*, *J. Clin. Investigation*, **30**, 353, 1951.

lost (thermolysis). Heat production is the result of the chemical processes of metabolism; therefore its regulation is sometimes called the *chemical regulation* of temperature. Heat is lost by physical processes, and its regulation is therefore called the *physical regulation* of temperature. The temperature level or homoiothermia (thermal homeostasis) is dependent on a well-controlled equilibrium between the chemical and physical processes of temperature regulation. This equilibrium can be represented (Fig. 220) by a balance, on one plate of which the processes of heat production are placed, and on the other those of heat loss. An increase in heat production tends to raise the temperature, while an increase in heat loss tends to lower it.

FIG. 220. Factors in the regulation of temperature. (*Du Bois, E., Lane Medical Lectures, Stanford University Press, Stanford, California, 1937.*)

THE CONTROL OF HEAT PRODUCTION

The amount of heat produced depends on the metabolic rate and therefore on the rate of exothermic chemical processes. Some of the intermediary exothermal chemical reactions are anoxybiotic (*i.e.*, not oxidations), but ultimately all the heat produced in the body comes from oxidations.

The following (in order of importance) are the principal mechanisms that regulate heat production (chemical regulation of temperature): (*a*) phasic muscular activity (exercise, shivering); (*b*) muscle tone; (*c*) the specific dynamic action of foodstuffs; (*d*) changes in the basal metabolic rate.

Glandular tissue, especially the liver and in second place the kidney, produces as much heat per unit weight as resting muscle or more, but the muscles are the most important factor in heat production. This is due to two reasons: (*a*) muscles form approximately one-half of the body mass; (*b*) muscular activity causes an increase in the metabolic rate, which can rise to from 4 to 10 times the resting level in strenuous exercise. Of the extra energy evolved in exercise, only 20 to 25 per cent is converted into mechanical energy; 75 to 80 per cent is given out as heat.

Muscle tone is a form of muscular activity that normally produces a large amount of heat and contributes to maintain a normal body temperature. It is increased by cold and emotion, even when there is no perceptible change. This can be demonstrated by registering the action potential of the muscle fibers; a considerable increase in asynchronous contractions

of motor units will thus be revealed.[1] Cold also stimulates voluntary movements in man and many animals, but the fundamental processes in the increase of heat production in response to cold are tremor and shivering. When a subject is placed in a cold environment there is no definite increase in heat production until shivering begins; then there is an immediate and considerable increase in oxygen consumption and heat production. This mechanism is much more important than any other in the response to cold.

FIG. 221. Effect of temperature on metabolism.

Phasic (movement, shivering) and tonic muscular contractions are controlled by the nervous system. Therefore if nervous activity is suppressed by anesthesia or the muscles are paralyzed with curare, heat lost by the body is not adequately replaced and the temperature falls. This fall is marked and rapid in animals intoxicated with curare.

Heat production is at a minimum in the majority of mammals when the outside temperature is 28 to 30°C. (82.4 to 86°F.). This is known as the critical temperature. At temperatures below 28°C., heat production increases so as to compensate the loss of heat, and at higher temperatures it also increases because there is a slight rise in body temperature, which increases the velocity of the chemical reactions in the organism according to van't Hoff's law. In man there is a fairly wide range of critical temperature, *i.e.*, environmental temperature conditions that have no influence on heat production. This range is given as 22 to 35°C. (71.6 to 95°F.) by Du Bois and 25.5 to 29°C. (77.5 to 83.4°F.) by Winslow and Herrington. In a damp atmosphere heat production and rectal temperature begin to increase when the temperature rises to 30 or 33°C. (84 or 91.4°F.). The increase in heat production caused by cold is considerable; it may be three and even four times the BMR. The highest metabolic rate provoked by cold in homoiothermic animals has been called "vertex metabolism."[1] In poikilothermic animals the metabolic rate increases as the temperature rises from 0 to 30°C. (32 to 84°F.) (Fig. 221).

Cold baths increase heat production because they cause movements and shivering. Hot baths increase heat production because they raise the body temperature. An increase in the temperature of the environment causes an increase in heat production only when it also causes a rise in body temperature of a few tenths of a degree centigrade.

The BMR increases in some species (*e.g.*, the rat) after they have lived for some time in a cold environment. This response is accompanied by changes in the liver, thyroid, adrenals, and other organs. When these animals are again placed in a warm environment (*e.g.*, if rats are taken from a room at 15°C. to another at 25°C.) the BMR returns to the lower level only after several days. Time is needed for the glandular changes to take place before the BMR is stabilized at a level corresponding to the environmental temperature.

The specific dynamic action of foodstuffs increases heat production. The effect of protein (30 per cent) is greater than that of fat (6 per cent) or carbohydrate (4 per cent). Probably this is the reason for the tendency to consume high-protein diets observed in countries with a

[1] BURTON, A. C., and D. W. BRONK, *Ann. Rev. Physiol.*, **1**, 109, 1939.

[1] GIAIA, M., *Compt. rend Soc. de biol.*, **90**, 1087, 1924; **92**, 364, 1925; **93**, 646, 1925.

cold climate. In such climates there is also a tendency to increase the amount of fat in the diet, especially if hard manual labor is performed, because fat has a high caloric value in small bulk. In several species, cold or a decrease in body temperature provokes a larger consumption of food, while heat or a rise in body temperature tends to diminish food consumption.

The increase in heat production and body temperature (fever) of infectious diseases will be considered later.

The Control of Heat Loss

All bodies when placed at a temperature below their own lose heat by conduction, radiation, convection, and vaporization of water. These processes also take place in homoiothermic animals, and the mechanism by which they are controlled is known as the physical regulation of temperature. In man the amount of heat lost depends on the body surface, the relation between the body temperature and the environmental temperature, and the humidity of the atmosphere. Most of the heat is lost by radiation, evaporation, and convection.

Conduction of heat. This consists in the transmission of heat from one molecule to another in a solid, liquid, or gaseous body. It depends on the conductivity of the body and the difference in temperature between the two points in question. Heat conductivity is measured by the amount of heat (in calories) transmitted in one second from one side to the opposite side of a cube of 1 cu. cm. of the substance, when there is a difference in temperature of 1°C. between the two sides. Differences in conductivity cause two bodies at the same temperature to appear to be at different temperatures when brought into contact with the skin; *e.g.*, wood (a poor heat conductor) appears to be warmer than metal (a good heat conductor) at the same temperature, because more heat is lost from the skin to the latter than to the former. Water is a relatively poor heat conductor, and air a very poor one. Man loses very little heat by conduction when in air; he loses more when he is submerged in water, but a larger amount may be lost by placing the body in contact with metals or a cold floor.

The specific heat of the body. The specific heat of pure water is the unit of specific heat, *i.e.*, the amount of heat needed to raise 1 kg. of water 1°C. is 1 kg.-cal. The specific heat of the human body is 0.83. A subject weighing 70 kg. with an average body temperature of 37°C. holds 2150 kg.-cal. ($70 \times 37 \times 0.83 = 2150$). The large proportion of water in the organism causes thermal changes to proceed at a low rate. Therefore, even if man had no mechanisms of temperature regulation, a large amount of heat would have to be gained or lost to cause a rise or a fall of 1°C. in temperature. In the example just given it would have to be 58 kg.-cal. ($70 \times 1 \times 0.83 = 58.1$). Water has a very high specific heat compared to that of air, which is from 0.169 to 0.237; moreover, it is a better heat conductor. Therefore a subject submerged in water at 0 to 5°C. loses much more heat, suffers more from cold, and dies sooner than when exposed to air at the same temperature.

Radiation. Radiant heat is transmitted by electromagnetic waves at the speed of light (300,000 km. per sec.). These waves show the same phenomena as light, *i.e.*, interference, reflection, refraction, and polarization. The heat from the sun is transmitted to the earth by these waves, which travel through the interstellar space and the air of the atmosphere. Radiant energy is converted into heat when it comes into contact with a material surface. If this surface absorbs 100 per cent of the impinging radiant energy it is known as a perfectly opaque body. Air absorbs very little radiant energy; water (*e.g.*, humidity in the air) absorbs much more. The human skin absorbs 97 per cent of the irradiation (infrared rays) that falls on it (Hardy and Muschenheim). It emits radiations of wavelengths varying from 50,000 to 200,000 A, with a maximum at 90,000 A (Hardy). These wavelengths are so far from those of visible light that the color of the skin (white or black) has no importance in this respect.

Radiation is proportional to the surface of the radiant body, to its power of emission, and to the difference between the absolute temperature of the radiant body raised to the fourth power and the absolute temperature of the coldest body receiving the radiation also raised to the fourth power.

The human body does not radiate heat to the surrounding air, but through the air to cold objects in the environment; *e.g.*, the walls, floor, ceiling, and furniture in a room, and through the windows to objects outside such as snow or ice. The essential factor in radiation is the difference in temperature between the bodies.

Thus if a subject is placed between a lighted stove and a cold wall, he will receive radiation from the hot stove and emit radiation to the cold wall. A perfect heating system heats the walls, ceiling, and floor of a room to the desired temperature.

Only 70 to 80 per cent of the body surface radiates heat, because certain parts, such as the axillae, the medial aspects of the thighs, etc., radiate to other parts of the body and not to objects in the environment. When many people are in the same room they irradiate each other, their body temperature rises, and discomfort is felt. "Vitiation" of the air has nothing to do with this process, but the increase in humidity of the atmosphere interferes with another mechanism of heat loss, *i.e.*, vaporization.

Convection. This consists in the transmission of heat in currents of liquids or gases resulting from differences in density or other causes. The air surrounding a radiant body is heated and rises; colder air replaces it, and thus currents are formed which carry away heat from the radiant body. The air between the layers of clothes, and between the meshes of the cloth out of which these are made, remains stagnant or circulates very slowly, so there is no loss of body heat by convection, and as air is a bad heat conductor it acts as an even more efficient insulator than clothes. More heat is lost by convection if air currents are artificially made by fans; loss of heat by evaporation can also be increased in this way. Convection increases with the speed of the air current up to 125 km. (78 miles) per hr.; a further increase in speed does not increase convection.

Evaporation. One liter of sweat absorbs 580 kg.-cal. on evaporation (pure water absorbs 539 kg.-cal.). Water evaporates on the skin and lung surfaces in direct proportion to the difference in temperature between the body and the environment and inversely to the percentage of humidity in the air. At temperatures between 18 and 30°C. (64 and 86°F.) the organism loses by evaporation from 22 to 27 per cent of the total heat loss. There are two types of evaporation on the skin, insensible perspiration and sweating.

Insensible perspiration consists in the evaporation of such a small amount of water that it vaporizes as soon as it is excreted. An adult man with a normal BMR loses 30 gm. of water per hour by this means; if the BMR rises, insensible perspiration increases and vice versa. Subjects without sweat glands lose approximately 18 gm. of water per hour by insensible perspiration.[1] Sanctorius (1614) was the first to measure insensible perspiration by placing the subject on a balance and noting the loss of weight, which is due mostly to the evaporation of water and only in small part to the difference in weight between CO_2 lost and oxygen absorbed.

Evaporation does not increase noticeably with an increase in environmental temperature up to 28 or 30°C. (81.4 or 86°F.). At this temperature there is a sudden increase due to sweat secretion, and at 35°C. (95°F.) nearly all the heat is eliminated by evaporation. At 37°C. (98.6°F.) heat cannot be eliminated from the organism by radiation; therefore evaporation accounts for all the heat lost. Heat is absorbed by the body from the environment when the outside temperature is above 37°C., but it is again eliminated by evaporation of sweat.

An optimum sensation of well-being is experienced with an environmental temperature around 20°C. (68°F.) and a relative humidity of 50 to 60 per cent. Evaporation diminishes as the humidity of the atmosphere increases. With a temperature of 37°C. and 100 per cent humidity, it would be impossible for the organism to lose heat either by radiation or evaporation; in this case the temperature of the skin rises above that of the environment and heat is thus eliminated by radiation but mainly by vaporization. Heat is tolerated better in a dry atmosphere than in a damp one; a subject in a warm and damp environment suffers discomfort, his temperature rises, and more serious effects may occur if the heat elimination is insufficient. This condition is observed in damp, hot valleys, near rivers, and in badly ventilated mines and workshops.

Man can withstand high temperatures in a dry atmosphere; *e.g.*, a temperature of 127°C. (260.6°F.) can be tolerated for 8 min. (Blagden), although this temperature is high enough to grill a steak or to bake bread or apples. In these conditions the organism must evaporate large quantities of sweat. Therefore an equivalent amount of water must be drunk and the body must be half naked or lightly clothed. If the atmosphere is saturated with humidity, a temperature of 48 or 50°C. (120°F.) cannot be tolerated for more than a few minutes. Subjects with congenital absence of the sweat glands

[1] RICHARDSON, H. B., *J. Biol. Chem.*, **67**, 397, 1926.

suffer considerably from heat in summer or when performing exercise; hyperthermia is easily provoked in them and is sometimes dangerous.

Heat is lost through the lungs, mostly by evaporation of water, and partly by convection. In certain animals (*e.g.*, the dog, which has sweat glands only on the pads of its paws) a rise in the temperature of the environment, or an increase in heat production due to strenuous exercise, causes what is known as thermic polypnea. The animal stretches itself against the ground (to increase the loss of heat by conduction), its breathing becomes rapid and shallow, and its tongue hangs out. The tongue is kept damp by an increased secretion of the salivary and buccal glands. Heat is thus lost by vaporization of water on the surface of the tongue. If tracheotomy is performed, thermic polypnea is much less efficient as a cooling mechanism. Polypnea is governed by impulses discharged from the nerve centers, owing to an increase in the temperature of the blood, and by reflexes arising in the skin.

Relative importance of the different ways of losing heat. The three principal ways of losing heat account for the elimination of 90 to 95 per cent of the total heat lost: (*a*) radiation, 55 per cent; (*b*) convection and conduction, 15 per cent; (*c*) evaporation through the skin and lung, 22 to 27 per cent. The remaining 5 to 10 per cent is used as follows: (*a*) about 60 kg.-cal. is used in warming the inspired air and is eliminated in the expired air: (*b*) 100 kg.-cal. is taken up by the elimination of CO_2 from the lungs; (*c*) 30 kg.-cal. is used to warm food and water ingested and is afterward eliminated in the feces and urine. Table 67, made up of data taken from Rubner, Martin, Du Bois, and others, summarizes the relative importance of the different ways of eliminating heat.

Table 67. Heat Elimination

Manner of elimination	*Daily heat loss*	
	kg.-cal.	%
Radiation, convection, and conduction	2100	68
Evaporation of water (skin and lungs)	810	26
Heating of inspired air	60	1.9
Elimination of CO_2 by the lungs	100	3.2
Urine and feces	30	0.9
Total	3100	100

There is not much change in the way heat is eliminated while the outside temperature varies from 15 to 28°C. (59 to 81.4°F.). Strenuous physical exercise produces a considerable change in this picture. Thus Du Bois and Hardy report an observation in which a subject at rest eliminated 66 per cent of the total heat by radiation, 15 per cent by convection, and 19 per cent by evaporation. Strenuous exercise caused heat production and elimination to be multiplied by 5; the amount of heat lost by radiation did not change, but heat loss by evaporation increased to such an extent that 75 per cent of the total heat was lost in this manner, and only 12 per cent by radiation and 13 per cent by convection. Rectal temperature rose from 37.5 to 38.8°C., and the temperature of the skin fell by 2°C., owing to the evaporation of sweat.

Relative importance of chemical and physical mechanisms for the regulation of body temperature. Body heat is lost by radiation and convection at low temperatures (up to 25.5°C., *i.e.*, 77.5°F.). Heat loss by evaporation increases rapidly from 29°C. (83.4°F.) up, and above 35°C. (95°F.) evaporation accounts for almost all the heat lost. In many species heat production (metabolic rate) increases at low temperatures and from 30°C. (85°F.) to 35°C. (95°F.) the metabolic rate may also increase.

When the environmental temperature falls, women lose about 10 per cent less heat than men. Perhaps this is due to the fact that they usually have more subcutaneous fat than men. The increase in the metabolic rate caused by low temperatures is also greater in women than in men. Moreover, women perspire less than men at the same environmental temperature, and their metabolic rate decreases at temperatures between 30 and 35°C. (Hardy and Du Bois).

Anatomic and physiologic factors in physical regulation of temperature. Several factors condition the loss of heat. Hair[1] and feathers in animals and clothes in man play an important part in controlling the loss of heat. They not only are poor heat conductors, but also enclose air, which is an even less efficient conductor of heat. When considering clothes as thermic

[1] If an animal is shaved or varnished it loses great quantities of heat and soon dies if it is placed in a cold environment.

insulators, the following factors should be taken into account: (*a*) area of the body covered; (*b*) nature, thickness, and color of the stuff used; (*c*) numbers of layers of clothes, because air between the layers is a very efficient insulator.

The skin and the poorly vascularized subcutaneous adipose tissue are poor heat conductors. Cutaneous circulation varies considerably in different circumstances. It plays an important part in the control of heat loss, because body heat is carried by the blood to the skin, where it can be eliminated. At an environmental temperature of 34°C. (93.2°F.) the skin may receive 12 per cent of the blood output of the heart. There is readjustment of the circulation in these conditions; when the skin vessels dilate, there is vasoconstriction of the splanchnic area and contraction of the spleen. The blood vessels of the skin are very sensitive to the effects of temperature. Heat causes cutaneous vasodilatation by (*a*) a direct action on the vasomotor centers; (*b*) vasomotor reflexes; (*c*) axon reflexes; (*d*) a direct effect on the blood vessels. Cold, on the contrary, causes cutaneous vasoconstriction. The skin becomes cold, the circulation is sluggish, and the blood in the cutaneous vessels becomes cyanotic. Nevertheless if the skin is rubbed with snow or an arm is submerged in freezing water, intense but transitory vasodilatation is observed, the skin becomes red and warm, and there is hyperalgesia.

An increase in the temperature of the environment produces the following effects: (*a*) the skin becomes red and warm; (*b*) the circulating blood volume increases, because water is reabsorbed from the tissue fluids in the skin, muscles, and liver, and because the spleen contracts and ejects stored blood; (*c*) if the temperature rises to 30°C. (86°F.) or more, the heart rate, and sometimes the systolic volume, increases; (*d*) respiratory frequency increases; (*e*) if there is hyperventilation, alkalosis is provoked, with transitory increase in the pH and decrease of ammonia in urine; (*f*) if the environmental temperature rises above 30 or 35°C. (86° or 95°F.), there is a slight increase in rectal temperature and heat production.

Man protects himself from heat by (*a*) cooling the atmosphere, or increasing the temperature gradient by cooling the surrounding structures; (*b*) decreasing the area of skin covered by clothes, and the thickness of the latter, choosing them of white or light color; (*c*) increasing the surface of the skin that can be used for radiation and convection; (*d*) cutaneous vasodilatation, which increases heat loss by radiation and evaporation; (*e*) increasing the evaporation of sweat by producing air currents with fans, to increase convection and evaporation; (*f*) diminishing the intake of food, especially of protein and fat. The heat of the sun is tolerated better by subjects covered by light clothes and a hat than by naked subjects. Clothes protect the body not only from cold, but also from excessive heat.

Man protects himself from cold by (*a*) covering a larger area of the body surface with several layers of thicker clothes, preferably of dark colors; (*b*) cutaneous vasoconstriction; (*c*) diminishing the surface of irradiation; (*d*) increasing heat production by exercise, high muscle tonus, and shivering; (*e*) increasing the caloric intake, especially in the form of protein and fat; (*f*) heating the environment (stoves, central heating, etc.).

Sweat Secretion

Sweat is a fluid excreted by the sweat glands in the skin. The sweat glands are highly developed in man and in the horse, but there are few or none in other species; thus the dog and cat have sweat glands only on the pads of their paws. Nevertheless the cat has been used quite as frequently as man or the horse in work done on the physiology of sweat secretion.

Sweat should be distinguished from insensible perspiration, which has been given this name because it is an imperceptible loss of water through the skin, even in subjects in whom the sweat glands are congenitally absent. At the rate of approximately 30 gm. per hr., between 600 and 800 gm. of water is evaporated by insensible perspiration every 24 hr., and this evaporation accounts for one-quarter of the heat lost by a resting subject. Insensible perspiration is greatest on the palms of the hands and the soles of the feet; next on the back of the hands, the neck, and the face; and least on the rest of the body surface.

Sweat glands. There are two types of sweat glands, according to Kuno:

1. *Eccrine glands*, distributed over the whole surface of the body, secrete a dilute fluid. There are from 2,000,000 to 3,000,000 of these glands. Placed on end, one after the other, they would extend over 2½ miles, and their aggregate mass is equivalent to

half that of one kidney. They are more numerous on the palms of the hands, the soles of the feet, the neck, and the trunk.
2. *Apocrine glands* are larger. They are found in the axillae, around the nipples, and on the mons veneris and labia. Their secretion has a characteristic odor.

The secretion of sweat is not a simple filtration of blood plasma through the glandular epithelium, because the pressure of sweat secretion (measured by a manometer occluding the glandular duct) can rise above the blood pressure (*e.g.*, to 250 mm. Hg). When the glands are active, circulatory, histologic, physical, and chemical phenomena typical of true secretion are observed. There is usually marked vasodilatation when the glands are secreting, but it is possible to dissociate the glandular and vascular responses. Thus secretion can be provoked by nerve stimulation in the sweat glands of a limb in which the circulation has been suppressed by ligatures or of one that has just been amputated. In certain circumstances sweat secretion occurs at the same time as vasoconstriction; in the cold sweat of intense emotion (anguish, fear) or in anoxia, the skin is pale. This type of sweating is more marked in the palms of the hands, the soles of the feet, and the forehead.

Sweat. The composition of sweat varies according to the place on the body surface from which it is collected, and the cause that has provoked its secretion. Sweat is a dilute watery solution; its specific gravity is 1.002 to 1.003, seldom higher. Its chemical composition[1] has been determined under different conditions (Table 68). The pH varies from 5 to 7.5. Usually it is on the acid side, especially when there is an abundant secretion due to heat. Sweat provoked by exercise is less acid than "heat" sweat; it has more ash and a higher concentration of organic substances. Chloride is eliminated as NaCl, in concentrations of 0.15 to 0.5 per cent. Nonprotein nitrogen eliminated by the sweat amounts to 0.071 gm. per day; it increases when there is retention of urea or nonprotein nitrogen in the blood.

A rise in body or environmental temperature causes the secretion of sweat to increase. Maximum rates of 1 and 2 liters per hr. have been observed, and strenuous exercise in the sun may cause the loss of 4 liters in 1 hr. Heavy physical labor in a hot, dry atmosphere may cause the secretion of 8 to 11 liters of sweat in 5 to 8 hr. of activity. In one subject the sweat eliminated in the course of 14 days amounted to 79 liters—more than the weight of the body. The excretion of such large quantities of sweat causes the loss of important amounts of water and salt; up to 10 gm. of NaCl can be lost in this way in 3 hr. If this happens the NaCl concentration in blood plasma and tissue fluids decreases, and the subject suffers from cramps, general malaise, insomnia, and in some cases fever. These disturbances are cured or prevented by drinking saline solution (0.25 to 1 per cent NaCl) instead of water, or by taking NaCl in tablets.

Table 68. Properties and Composition of Sweat

	Range	*Usual value*
Density	1.001–1.006	1.003
Freezing point, °C.	−0.13–−0.54	−0.24
pH	5–7.5	5.2 (heat) 6.6 (exercise)
Water, %	99.2–99.6	
Solids, mg. %	258–890	680
Organic substances, mg. %	30–290	
Ash, mg. %	144–566	
Chlorine, mg. %	70–346	180
Potassium, mg. %	....	17
Sodium, mg. %	75–250	150
Sulfate, mg. %	4–6	5
Total N, mg. %	34–160	
Nonprotein N, mg. %	23–94	42
Urea N, mg. %	17–38	26

Note: Lactic, formic, acetic, propionic, butyric, capronic, caprylic, and citric acids are also found.

The innervation of the sweat glands. The sweat glands are innervated by cholinergic fibers of the sympathetic nervous system but there are also some adrenergic fibers. Sweat secretion is obtained by stimulating the sympathetic, or nerves which contain sympathetic fibers (*e.g.*, the sciatic nerve), the dorsal or cervical spinal cord, the medulla, the hypothalamus, and to a lesser degree certain cortical areas. Stimulation of the sympathetic fibers that cause sweat secretion liberates acetylcholine, which passes into the sweat. Acetylcholine injected into the arm artery provokes marked

[1] Talbert, G. A., *J. A. M. A.*, **87**, 1829, 1926; *Am. J. Physiol.*, **81**, 74 and 81, 1927; Whitehouse, *Proc. Roy. Soc., London, s.B.*, **117**, 139, 1935; Couraud, J., thesis, Bordeaux, 1935.

vasodilatation and sweat secretion on the palm of the hand. Acetylcholine and pilocarpine stimulate the secretion of sweat; atropine inhibits it. Acetylcholine and pilocarpine act directly on the gland, even after it has been denervated, although several weeks or months after denervation the response is less marked.

An intravenous injection of adrenaline does not provoke sweat secretion, but an intradermal one provokes local sweat secretion in 84 per cent of the cases. Noradrenaline has the same effect in 74 per cent of the individuals. Drugs that inhibit the sympathetic, *e.g.*, dibenamine, suppress this local action of adrenaline.

The sympathetic nervous system is of fundamental importance in the regulation of heat loss by evaporation of sweat. If the temperature rises, sweat is secreted by all the sweat glands, except those in a denervated area. The nerve centers are very sensitive to variations in the temperature of the blood; thus, if the blood in the carotid arteries is warmed, sweat secretion is stimulated, although there has been no increase in rectal temperature.

Sweat secretion is stimulated by (*a*) an increase in the temperature of the environment, heavy clothing, hot baths, or high-frequency currents (diathermy); (*b*) physical exercise, which causes body temperature to rise and impulses to be discharged from the nerve centers; (*c*) psychic factors, as intense emotional states, fear, etc.; (*d*) stimulation of nerve centers by asphyxia or other causes; (*e*) sleep; (*f*) reflexes, *e.g.*, local warming of a limb, centripetal stimulation of the sciatic or the splanchnic nerves in man, certain gustatory stimuli, etc.

The principal nerve centers that control sweat secretion are situated in the hypothalamus (Karplus and Kreidl). Sweat secretion is provoked by stimulating the hypothalamus electrically or by heating it locally by means of a thermoelectric apparatus or of tubes through which hot water circulates; also by injecting pilocarpine into the lateral ventricle (Cushing). The spinal centers are also sensitive to heat, although not so much as the hypothalamus. For this reason, transverse section of the spinal cord does not suppress sweat secretion below the section, and all the glands in the body respond to adequate stimuli. Section of the dorsal (afferent) roots does not suppress sweat secretion, because the secretory fibers leave the spinal cord in the ventral roots. Complete extirpation of the spinal cord below the first thoracic segment, in a patient with a progressive glioblastoma, suppressed sweat secretion provoked by heating except in the head and neck, although the sweat glands responded to direct chemical stimulation with furmethide.[1]

Sweat secretion can be inhibited by local application of cold on the skin, *e.g.*, submerging a limb in a cold bath, or for a short time by drinking iced water. When a subject lies on one side of the body, sweat secretion diminishes on that side and increases on the opposite side.

Sweat secretion has been studied mainly in man, horse, and cat (the pads of the paws). Insensible perspiration is measured by recording the loss of weight of the subject, or by measuring the increase in weight of absorbent substances placed over a small area of skin and hermetically covered. Sweat is collected by means of absorbent pads or in a rubber glove. If paper impregnated with silver nitrate is placed over the skin, when the sweat secretion begins, the openings of the sweat glands leave an impression on the paper (Aubert). If the skin is dusted with talcum powder containing tartaric acid and blue litmus, over the area where sweating occurs the powder takes on a red color owing to the acidity of the sweat. Sweat secretion is provoked experimentally by heat, exercise, pilocarpine, or stimulation of nerves (sympathetic, sciatic, etc).

The Nervous System in the Regulation of Body Temperature

The mechanisms that regulate body temperature are integrated by the nervous system. The young of homoiothermic species that are born mature, *i.e.*, with a well-developed nervous system, having their eyes open, and walking and feeding themselves (*e.g.*, horses and cows), resist cold fairly well. On the other hand, the young of species that are born with an imperfectly developed nervous system, with their eyes closed, or with incomplete locomotion (*e.g.*, man, rat, pigeons), do not resist cold (Milne, Edwards, Pembrey).

Peripheral nerves. The following peripheral nerves take part in the regulation of temperature: (*a*) motor nerves, which stimulate movements and tone in muscles (there is a rapid fall in temperature in animals intoxicated with curare, which paralyzes the muscles, and in animals in

[1] MacCarty, C. S., G. M. Roth, and G. J. Thompson, *Proc. Staff Meet., Mayo Clin.*, 26, 113, 1951.

which the spinal motor centers have been destroyed); (*b*) vasomotor nerves, which constrict and dilate cutaneous nerves; (*c*) the nerves of the sweat glands. After bilateral removal of the whole sympathetic chain in cats, the animals have a normal temperature as long as they are kept in a warm room, but the response to cold is disturbed. There is no erection of hairs or cutaneous vasoconstriction, and in spite of their shivering at higher temperatures than the controls, their body temperature falls more easily than in normal animals. In a hot atmosphere their body temperature rises sooner than that of the controls.[1]

Cerebral cortex. The decorticate dog maintains a normal temperature and responds normally to heat and cold. Nevertheless shivering occurs sooner than in the controls when it is exposed to cold, and the body temperature falls easily; when it is exposed to heat, thermal polypnea appears later than in the controls. Removal of the frontal lobes provokes chronic disturbances in the vasomotor reactions of the skin in dogs and monkeys. Probably the cortex is more important in body-temperature regulation in man than in other species.[2]

Hypothalamus. The principal centers that integrate body-temperature regulation are situated in the hypothalamus. Section of the brain rostral to the hypothalamus provokes little or no change in body temperature, even when the striate (which was once considered the main center of temperature regulation) is removed. A transverse section of the brain immediately caudal to the hypothalamus causes serious disturbances in the regulation of body temperature.[3] Animals thus treated resemble poikilothermic animals, because their temperature falls rapidly if they are exposed to a temperature of 20°C., and it is necessary to keep them in a room at 30 to 35°C. to keep them alive. The hypothalamus is sensitive to temperature changes and provokes responses adequate to restore the normal body temperature. Thus heating the hypothalamus causes vasodilatation, sweating, and hyperpnea (Barbour),[1] while cooling it causes vasoconstriction and tremor.

Two areas have been differentiated by experiments in the hypothalamus. The rostral part of the hypothalamus integrates responses that cause loss of heat and prevent hyperthermia. This area is situated above and rostrally to the optic chiasma, under the anterior commissure (supraoptic and preoptic nuclei). Warming this area provokes vasodilatation, sweating, and hyperpnea in the cat. Destruction of this area brings about the loss of normal response to heat; there is no hyperpnea, and hyperthermia occurs when the animals are placed in a warm environment. Fatal hyperthermia has been observed in animals and man in cases of destructive lesions in this area.

The caudal part of the hypothalamus integrates the physiologic response to cold, which prevents the fall of body temperature. Bilateral destructive lesions of the caudal part of the lateral nuclei of the hypothalamus, provoked experimentally or observed in patients, provoke a decrease in BMR and subnormal temperature; on exposure to cold, vasoconstriction occurs, but there is no shivering. There are sympathetic centers in this area, and probably there are also centers that exert some influence on the thermal tonic contractions of striated muscles, perhaps through tectospinal and rubrospinal paths. The temperature-regulating impulses discharged from the hypothalamus stimulate medullary and spinal centers; they do not act directly on peripheral structures (see Chap. 84, Functional patterns integrated in the hypothalamus).

The spinal cord. A transverse section of the nerve centers anywhere between the superior colliculi and the lower segments of the cervical spinal cord converts a homoiothermic animal into a poikilothermic one, because of the separation of the spinal cord from the hypothalamic temperature-regulating centers.

In dogs, if the spinal cord is cut at the level of the first thoracic segment, heat regulation by sympathetic control (vasomotor nerves, etc.) is lost, but the dogs can maintain their body temperature in a cold environment, because the muscles of the neck and forelimb remain innervated; shivering limited to a restricted area can, therefore, be enough to maintain a normal

[1] Sawyer, M. E. M., and T. Schlossberg, *Am. J. Physiol.*, **104**, 172, 1933.

[2] Bard, P., and D. McK. Rioch, *Am. J. Physiol.*, **109**, 515, 1934; *Bull. Johns Hopkins Hosp.*, **60**, 73, 1937.

[3] Isenschmidt, R., and L. Krehl, *Arch. f. exper. Path. u. Pharmakol.*, **70**, 109, 1912; Isenschmidt, R., and W. Schnitzler, *Arch. f. exper. Path. u. Pharmakol.*, **76**, 202, 1914; Bazett, H. C., *et al.*, *Arch. Neurol. & Psychiat.*, **30**, 728, 1933; Ranson, S. W., and H. W. Magoun, *Ergebn. d. Physiol.*, **41**, 56, 1939; Blair, J. R., and A. D. Keller, *J. Neuropath. & Exper. Neurol.*, **5**, 240, 1946.

[1] Magoun, H. W., *et al.*, *J. Neurophysiol.*, **1**, 101, 1938.

body temperature. On the other hand, if the cervical and brachial plexuses are cut and the lumbosacral segments of the spinal cord are destroyed, the dogs cannot maintain a normal body temperature in a cold environment.[1] In man lesions of the cervical spinal cord usually provoke hyperthermia; complete section usually causes hypothermia,[2] but in some cases hyperthermia has been observed.[3]

Endocrine Regulation of Body Temperature

Thyroid. The thyroid gland shows signs of greater activity in winter than in summer. When an animal is taken from a warm to a cold environment, histologic signs of thyroid hyperactivity are observed. This does not take place if the hypophysis has been previously removed or the hypophyseal stalk has been cut.[4] It therefore seems that cold acts on the hypothalamus and provokes a discharge of impulses that increase the secretion of thyrotrophin from the anterior hypophysis, thus stimulating the thyroid.

Thyroidectomized animals have a cold skin, and frequently their temperature is subnormal. Their heat regulation is deficient; thus they react to the injection of pyrogenous vaccines with a less marked rise in body temperature than is seen in normal animals.

In hyperthyroidism the skin is usually warm, and the temperature is high but within the normal range. Hyperthermia is not usually observed because, although heat production is increased, heat loss increases sufficiently to maintain a normal temperature balance. Some hibernating animals can be awakened from the winter sleep by the injection of thyroxine, which increases their BMR and body temperature.

Adrenals. Injection of adrenaline or the sudden discharge of adrenaline from the adrenal medulla causes an increase in the metabolic rate and in some cases a rise in temperature. On cooling an animal, there is a discharge of adrenaline, which increases heat production (Cannon *et al.*). Adrenalectomized animals do not resist cold well (Cannon, Marval), but this is due mainly to corticoadrenal insufficiency, not to the lack of adrenaline secretion. Corticoadrenal treatment restores the normal response to cold in these animals.

Hypophysis. Anterior hypophyseal insufficiency causes a decrease in the resistance to cold and a definite fall in the BMR (25 to 40 per cent, according to the species). In this condition there is a certain degree of adrenal and thyroid insufficiency, which adds to the effect produced by the absence of hypophyseal hormones. The anterior hypophysis exerts a continuous indirect action on heat production by maintaining a normal level of thyroid and corticoadrenal functions.

FEVER

The rise in temperature observed in fever is due to a considerable increase in heat production, which is frequently doubled and sometimes reaches five times the BMR, as occurs in an attack of malaria or similar conditions. In the first stage there is shivering, cutaneous vasoconstriction, paleness, goose flesh, and a sensation of cold. Heat production is increased (shivering) and heat loss is decreased (cutaneous vasoconstriction); therefore the body temperature increases because more heat is produced than is eliminated. When the temperature reaches its peak, the skin becomes warm because of cutaneous vasodilatation and heat loss increases. If the fever recedes (crisis), there is muscular relaxation and abundant sweating. Heat loss then predominates over heat production and the temperature falls.

An increase of body temperature up to 40°C. (104°F.) is frequently observed, but a rise above this temperature seldom occurs, as if there were mechanisms that controlled these higher rises.

When there is continuous fever, heat control is adjusted to maintain a high temperature. The patient sweats and the cutaneous blood vessels dilate if he is covered to excess or submerged in a warm bath; on the contrary, he shivers, the cutaneous blood vessels contract, and he has goose flesh if he is submerged in a cold bath. Heat control is not so firm in these conditions as in normal subjects; *e.g.*, in a cold bath the patient's body temperature falls more easily than that of a healthy person. Also exercise easily causes an increase in body temperature in febrile subjects (tuberculosis, convalescence of typhoid fever, etc.).

The following signs and symptoms are present in fever: malaise, mental torpor, headache, depression, delirium, loss of appetite, constipation. They dis-

[1] Chatonnet, J., *J. de physiol et de path gén.*, **43**, 678, 1951.
[2] Holmes, G., *Brit. M. J.*, **2**, 815, 1915.
[3] Foerster, O., *Handb. Neurol.*, **6**, 1, 1936.
[4] Uotila, U. U., *Endocrinology*, **25**, 605, 1939.

appear when the temperature falls as a result of treatment with antipyretic drugs or cold baths.

The metabolic rate increases in fever in proportion to the increase in temperature; there is approximately a 13 per cent rise for every degree centigrade of temperature above the normal. The pulse rate increases. The blood pressure rises and remains high if the temperature increases suddenly, but returns to normal when there is a stable temperature, even if there is fever. Pulmonary ventilation increases as the temperature rises.

Protein breakdown increases in fever; therefore there is a greater excretion of nitrogen, urea, and creatinine. The minimum protein requirement to maintain nitrogen equilibrium is higher than in normal subjects, in spite of the fact that protein does not account for more than 20 per cent of the total calories spent by a fasting patient with fever. The sharp fall in temperature characteristic of the crisis is accompanied by polyuria and the excretion of a large quantity of urea. Fat and carbohydrate are consumed in larger amounts during fever. If sufficient carbohydrate is not ingested, there is a tendency to ketosis and acidosis.

During fever the blood becomes concentrated, blood chlorides decrease, and little or no chloride is excreted in the urine. Ingested chloride is retained in great part in the tissue fluids. When the temperature drops, there is polyuria and the retained chloride is excreted.

Fever is probably due to disturbances in the hypothalamic heat-regulating centers. Many antipyretic drugs produce their effect by acting on the hypothalamus.

Treatment by hyperthermia. Certain diseases (rheumatism, gonorrhea, syphilis, mental diseases, etc.) are treated by provoking hyperthermia in the patient. Hyperthermia can be produced by (*a*) infection with *Plasmodium malariae;* (*b*) bacterial vaccines; (*c*) high-frequency currents (diathermy); (*d*) specially constructed ovens. Temperatures of 39 and 40°C. (102 and 104°F.) and pulse rates of 160 or more are provoked by these means. A severe strain is therefore placed on the organism, and the patient should be carefully watched during the whole process.

ACCIDENTS DUE TO COLD AND HEAT

Exposure to cold after the response of the organism has been exhausted produces weakness, sleep, torpor, coma, and death. If the body temperature falls to between 20 and 24°C. (66 and 75.2°F.) the subject seldom recovers, but if the temperature has fallen only to 27 or 29°C. (80.6 or 84.2°F.) and the subject is adequately treated by being rapidly warmed in a hot bath, recovery frequently occurs. Local freezing of tissues causes disturbances in local circulation, with the eventual mortification and destruction of the tissues.

Excessive heat can overcome the resistance of the organism with serious and even fatal results (heat stroke). Sweating diminishes; the metabolic rate, temperature, and pulse rate increase; there is hyperpnea, exhaustion, malaise, and irritability or somnolence. The mechanisms that regulate body temperature do not function, and the nerve centers suffer from the increase in temperature. The respiratory center of mammals ceases to function soon after the temperature rises to 44 or 45°C. (111.3 or 113°F.). Death is due to circulatory collapse. Sometimes convulsions are observed; in other cases the subject becomes comatose in the final stages. A few cases of survival of patients having had a temperature of 45°C. (113°F.) have been reported.

TEMPERATURE REGULATION AND CLIMATE

Deficiency in heat loss produces discomfort. The principal factors which regulate heat loss through the skin and which depend on climate are (*a*) the difference in temperature between the skin and the environment; (*b*) the humidity of the atmosphere; (*c*) the movement of the air surrounding the individual. Hill[1] has suggested the use of the katathermometer. The thermometer is warmed and the speed at which it cools is determined; then it is again warmed, the bulb is surrounded by damp cotton, and the speed of cooling is again determined. The first curve corresponds to heat lost by radiation, and the second to that lost by evaporation. The influence of the humidity of the atmosphere on heat loss can thus be measured.

BIBLIOGRAPHY

ADOLPH, E. F., "Physiological Regulations," Cattell and Company, Incorporated, Lancaster, Pa., 1943.

ADOLPH, E. F., "Physiology of Man in the Desert," Interscience Publishers, Inc., New York, 1947.

[1] HILL, L., *et al.*, *Tr. Roy. Soc.*, *London*, *s.B.*, **207**, 133, 1916; Medical Research Council Rep. No. 73, 1923.

BAZETT, H. C., Physiological Responses to Heat, *Physiol. Rev.*, **7**, 531, 1927.

DILL, D. B., "Life, Heat and Altitude," Oxford, New York, 1938.

DU BOIS, E., "The Mechanism of Heat Loss and Temperature Regulation" (Lane Medical Lectures), Stanford University Press, Stanford, Calif., 1937.

DU BOIS, E., Heat Loss from the Human Body, *Bull. New York Acad. Med.*, **15**, 143, 1939.

DU BOIS, E., "Fever and the Regulation of Body Temperature," Am. Lect. Series, No. 13, Charles C Thomas, Springfield, Ill., 1948.

HARDY, J. D., Physiological Responses to Heat and Cold, *Ann. Rev. Physiol.*, **12**, 119, 1950.

KUNO, Y., "Physiology of Human Perspiration," J. & A. Churchill, London, 1935.

MACHLE, W., and T. F. HATCH, Heat, Man's Exchanges and Physiological Responses, *Physiol. Rev.*, **27**, 200, 1947.

NEWBURGH, L. H., "Physiology of Heat Regulation and the Science of Clothing," Saunders, Philadelphia, 1949.

RANSON, S. V., Regulation of Body Temperature, in The Hypothalamus, *Assoc. Res. Nerv. Mental Dis.*, **20**, 342, 1940.

SUNDSTROEM, E. S., The Physiological Effects of Tropical Climate, *Physiol. Rev.*, **7**, 320, 1927.

THAUER, R., Der Mechanismus der Warmregulation, *Ergebn. d. Physiol.*, **41**, 609, 1939.

WINSLOW, C.-E. A., and L. P. HERRINGTON, "Temperature and Human Life," Princeton University Press, Princeton, N. J., 1949.

CHAPTER 49

Vitamins

VITAMINS ARE ORGANIC substances which the tissues cannot synthesize in sufficient quantities by metabolic processes. They must be found in food, preformed or as "provitamins" which are converted into vitamins, *e.g.*, carotenes are transformed into vitamin A. Some vitamins are produced by bacteria in the digestive tract, *e.g.*, vitamin K, but not in sufficient amounts to satisfy the needs of the body. Only very small amounts of vitamins are needed for normal nutrition, but these are indispensable, otherwise a specific deficiency disease appears, *e.g.*, rickets, beriberi, scurvy, etc. Several vitamins form part of coenzymes in the enzymatic systems of fundamental metabolic processes.

The term "vitamin" was proposed in 1912 by Funk, who believed there was one amine indispensable for life. Later work showed that there were several indispensable substances and that most of them were not amines.

Long before vitamins were discovered, the disturbances caused by their deficiency had been described, and methods of curing these diseases by giving certain foods, which are now known to contain the missing vitamin, had been discovered. The effect of the complete lack of a vitamin is known as "avitaminosis" and that of an insufficient amount as "hypovitaminosis." In many cases weeks or even months pass before the signs of deficiency are evident, because the organism stores vitamins and these stores take time to be exhausted. In the first stages of the disease it is easy to obtain a complete cure very rapidly by the administration of the missing vitamin, but after a time it is more difficult to restore the normal condition. Sometimes a complete cure is not obtained, or the process may have become irreversible. Slight insufficiencies are not always easy to recognize, and this fact may have social consequences because the health of large numbers of the community may be impaired.

Vitamins are unstable; therefore they are easily destroyed by oxidation in the process of preparing foods and in stored foodstuffs. They are preserved by quickly drying the foods at low temperature and by storing them dry in an atmosphere free from oxygen. Many of the vitamins are destroyed by heat, especially in an alkaline medium. Food that contains them should therefore not be submitted to prolonged cooking at temperatures above 100°C.

Vitamins have not only been isolated in the pure state but also prepared by synthesis. Thus large quantities, easy to preserve, can be obtained at low cost for therapeutic, preventive, and experimental uses. This has made possible the observation of the effects on the organism of extraordinarily large amounts of vitamins. In some cases disturbances known as hypervitaminoses have been seen. Hypervitaminoses never occur spontaneously in man; they are diseases provoked in laboratory animals by repeated administration of excessively large doses of one or another of the vitamins. The condition may become severe and the animals may die if the treatment is sufficiently prolonged. In man a few cases of hypervitaminosis D have been reported as a result of excessive doses of the antirachitic vitamin taken over long periods without adequate supervision.

Vitamins are not only indispensable to higher animals; they are also necessary for growth and development of bacteria and yeasts. A vitamin that is essential for one species, however, is not necessarily indispensable to another. Some of the vitamins (thiamine, riboflavin, niacin) are synthesized by the bacteria in the intestine, which are thus a source of vitamins partially satisfying the requirements of the host. Certain

intestinal disinfectants, such as sulfamides or antibiotics, diminish the production of these vitamins by inhibiting bacterial growth. The same effect is observed in certain intestinal diseases.

The activity of vitamins and their concentration in foods is estimated by physical, chemical, or biological methods. The latter consist in the experimental production of the required avitaminosis in animals of a species susceptible to the deficiency, which are then treated with different foodstuffs with the object of observing the therapeutic effects. Recently microbiological methods, *i.e.*, the use of bacteria, have been more frequently employed in the assay of vitamins. Controls made by treating human cases should be used whenever this is possible.

Nomenclature. The term "provitamin" is used for substances that are not vitamins but can be converted into vitamins. When first discovered, the vitamins were distinguished from each other by letters, but later, as their chemical nature became known, the appropriate chemical names came into use. Nevertheless conventional names or letters are commonly used, because they are easy to remember. Up to now over twenty vitamins have been described. Many of them are indispensable for growth of bacteria or yeasts; here only those of importance in human nutrition will be considered. A primitive classification of vitamins divided them into fat-soluble and water-soluble vitamins, because they were first found dissolved in natural fats or extracted by water from different foodstuffs. Some of the vitamins form groups of chemically related substances, with similar biological activity.

Group A: A_1, axerophthol; antixerophthalmic.
A_2, similar to A_1.

Group B: B_1, thiamine or aneurine, antineuritic or antiberiberi.
B_2, riboflavin or lactoflavin.
B_6, pyridoxin, antidermatitis factor in the rat.
B_{12}, extrinsic antianemic factor.
Folic acid, pteroylglutamic acid, antianemic.
Nicotinic acid (niacin), nicotinic amide (niacinamide), P-P or pellagra-preventing factor.
Human requirements of some important vitamins are not well known, *i.e.*, pantothenic acid, biotin, inositol, paraminobenzoic acid (PABA), streptogenin.

Vitamin C: Ascorbic acid, antiscorbutic.

Group D: D_2, calciferol (ergocalciferol) derived from ergosterol by irradiation; antirachitic.
D_3, (cholecalciferol), D_4, D_5, and D_6, similar to D_2.

Group E: α, β, and γ tocopherols; antisterility factor; human requirement unknown.

Group K: K and K_2, coagulation vitamin; antihemorrhagic, stimulates formation of prothrombin.

Vitamin P: Citrin or hesperidin; associated with ascorbic acid in maintaining capillary-wall resistance.

The human requirement of vitamins increases in the following circumstances: (*a*) increase in metabolic rate, *e.g.*, manual labor, high BMR; (*b*) increase in catabolism, as in hyperthyroidism and diabetes; (*c*) pregnancy and lactation; (*d*) growth; (*e*) deficient absorption, *e.g.*, alcoholism; (*f*) certain chronic diseases, convalescence, and healing of wounds.

Ingestion, absorption, storage, or transformation of vitamins can be disturbed in several pathologic conditions. Ingestion and absorption are hindered in diseases of the mouth, stomach, and intestines. Absorption of water-soluble vitamins is disturbed by gastric achylia and alcoholism. Absorption of fat-soluble vitamins is disturbed when there is obstruction of the bile ducts, or functional insufficiency of the intestinal mucosa, or commonly when cathartics such as mineral oil are administered over a period of time. Storage and transformation of vitamins A, D, K, and B are deficient when the liver cells are damaged. In these cases the deficiency is said to be secondary, since it is not due to a primary deficiency of the vitamin in the diet.

Antivitamins. Certain enzymes in foods destroy some of the vitamins. There are also substances with a chemical structure similar to the vitamins which cause avitaminosis by the mechanism of substrate competition,[1] similar to the mechanism by which carbon monoxide competes with oxygen for hemoglobin. These so-called "antivitamins" will be considered when dealing with each vitamin in particular.

[1] Woolley, D. W., *Physiol. Rev.*, **27**, 308, 1947; *Ann. New York Acad. Sc.*, **52**, 1197, 1950.

VITAMIN A

Vitamin A is known as the "antixerophthalmic" vitamin, because its deficiency causes a disease in the eye called xerophthalmia. Its chemical structure is known and it has been prepared by synthesis, but there is at least one other substance (vitamin A_2) and perhaps several, of unknown chemical structure, which have the activity of vitamin A.

Avitaminosis and Hypovitaminosis

Vitamin A maintains epithelial tissues in normal condition, thus protecting them from physical or bacterial aggression. Deficiency of vitamin A causes atrophy of epithelial tissues, followed by proliferation of the basal cells, with abnormal formation of large numbers of cornified cells (Wolbach).

Vitamin A deficiency in man. The following signs and symptoms are observed in man when there is deficiency of vitamin A:

1. *Xerophthalmia.* There is thickening of the con junctiva, which loses its transparency; lachrymal secretion diminishes, the conjunctiva becomes dry, and the conjunctival folliculi are hypertrophied, especially on the lower lid (follicular conjunctivitis).
2. *Keratomalacia.* In advanced stages of severe cases the cornea is thickened and ulcerated, and sometimes the eyeball is perforated.
3. *Papular eruption in the hair follicles with keratosis* of the area around them gives the skin the aspect of goose flesh. This is especially marked on the arms and legs.
4. *Cutaneous xerosis, i.e.,* dry, scaly skin, with or without follicular keratosis.
5. *Night blindness* (nyctalopia), consisting in failure of vision in twilight or at night, is frequently found in undernourished communities. It is due to vitamin A deficiency. The vitamin bound to a protein forms the visual purple (rhodopsin) in the retina. Exposure to light converts rhodopsin into retinene or xanthopsin (visual yellow), which is an aldehyde of vitamin A bound to a protein. Light acting on visual purple stimulates the optic nerve endings and sends impulses along the nerves. In the dark, retinene is reconverted into visual purple. If there is vitamin A deficiency, the process of visual purple regeneration is retarded. Administration of vitamin A rapidly restores the normal condition (see Chap. 78). Hypovitaminosis A can be detected by measuring the time of adaptation to darkness on passing from intense to dim light; this time is prolonged in cases of vitamin A deficiency. The test has no significance in isolated cases, but it is useful in the statistical determination of vitamin A deficiency in a community, especially when treatment with vi amin A accelerates dark adaptation, while individuals not treated show no change.
6. *Susceptibility to infections.* The epithelia of the digestive, respiratory, and genitourinary tracts become dry and cornified and are easily infected. Infections are frequently observed in the middle ear, eyes, nasal sinuses, bronchi, and the genitourinary tract. Treatment with vitamin A rapidly clears up the infection, a fact that has made some workers refer to this vitamin as the anti-infectious vitamin. If there is no deficiency, vitamin A even in large doses has no effect on the course of infections.

Experimental avitaminosis A. The following symptoms are observed in laboratory animals fed on a diet free from vitamin A:

1. *Loss of weight,* although the bones do not cease to grow.
2. *Ophthalmia.* The conjunctiva is dry and thickened and later becomes infected. The eyelids are red and swollen.
3. *Cornification of epithelia* of the respiratory, digestive, and urinary tracts, the vagina, and numerous glandular ducts. The epithelial linings are formed by flat, stratified cells, which become cornified, dry, and desquamatory.
4. *Suscebtibility to infections.*
5. *Nephrolithiasis.* Renal stones are frequently formed by desquamation of the keratinized epithelium of the renal pelvis.
6. *Degenerative changes in spinal and cranial nerves,* caused by compression of these nerves at their point of emergence from the cranium or spinal cord by thickening of the bone and narrowing of the foramina through which the nerves pass. This is due to excessive activity of the osteoblasts and insufficient activity of the osteoclasts.[1] There is insufficient evidence of its occurrence in man.

[1] Mellanby, E., *Proc. Roy. Soc., London, s.B.*, **132**, 28, 1944.

Hypovitaminosis A is diagnosed either by measuring adaptation to darkness (see above) or by estimating the concentration of vitamin A in the blood. It is treated by giving 25,000 international units (IU) of vitamin A daily.

Hypervitaminosis A has seldom been observed in man. It has been reported in patients treated with large doses of this vitamin and in persons who had eaten great quantities of polar-bear liver. Animals fed on concentrated liver oils have anorexia, loss of hair, weakness or paralysis, multiple fractures, hemorrhages, emaciation, and fatty degeneration of the liver (Takahashi, Collazo, *et al.*).

Sources of Vitamin A

Vitamin A is produced in the organism by the conversion of provitamins found in vegetables. These precursors are plant pigments called "carotenoids," which are found in green plants or in animals that have eaten these plants. In the green or yellow parts of plants that are vitamin precursors, over ten carotenoids have so far been found. The most significant of these are α, β, and γ (especially β) carotene. These pigments are absorbed in the intestine and are converted into vitamin A. They are stored mainly in the liver together with carotenes. Vitamin A is excreted into milk. Milk and butter are important sources of this vitamin in the human diet; therefore when butter is replaced by margarine, the latter should be reinforced by the addition of vitamin A. Cod-liver oil contains 600 IU/gm.; other fish-liver oils have even larger quantities, *e.g.*, shark 30,000 and halibut 60,000. There is no vitamin or provitamin A in vegetable oils, except in coconut oil and in corn oil. Cereals, with the exception of yellow corn, have little or no vitamin A. There is none in potatoes, but it is found in sweet potatoes. Green and yellow vegetables, such as lettuce, water cress, runner beans, peas, and carrots; fruits such as melons and bananas; and alfalfa contain carotenes that are vitamin A precursors. The vitamin A content of certain foods is given in Table 69.

Table 69. Vitamin Content in Foodstuffs

Food, 100 *gm. or* 100 *cc.*	*Vitamin A, IU*	*Thiamine, mg.*	*Riboflavin, mg.*	*Niacin, mg.*	*Ascorbic acid, mg.*
Butter	3,300	Trace	0.01	0.1	0
Cheese (Cheddar)	1,740	0.04	0.50	(0.2)	(0)
Egg (whole, fresh)	1,140	0.12	0.34	0.1	0
Milk (whole, fresh)	(160)*	0.04	0.17	0.1	1
Bacon (medium fat)	(0)	(0.42)	(0.10)	(2.1)	0
Beef (medium, side)	(0)	0.11	0.14	4.7	0
Liver (fresh)	19,200	0.27	2.80	16.1	31
Cod	......	0.04	0.05	2.3	2
Chicken, roasters	Trace	0.11	0.18	8.6	0
Apples	90	0.04	0.02	0.2	5
Lemons	0	0.04	Trace	0.1	45
Oranges	(190)	0.08	0.03	0.2	49
Tomatoes (fresh)	1,100	0.06	0.04	0.6	23
Tomatoes (canned)	1,050	0.05	0.03	0.7	16
Cabbage (fresh)	80	0.07	0.06	0.3	52
Carrots (fresh)	12,000	0.07	0.06	0.5	6
Peas, green (fresh)	680	0.36	0.18	2.1	26
Peas, green (canned)	540	0.11	0.06	0.9	8
Spinach (fresh)	9,420	0.12	0.24	0.7	59
Flour, wheat, patent	(0)	0.07	0.03	0.8	0
Flour, whole-wheat	(0)	0.56	0.12	5.6	0
Oatmeal	(0)	0.55	0.14	1.1	0
Rice, white	0	0.05	0.03	1.4	0
Yeast, dried brewer's	(0)	9.69	5.45	36.2	(0)

Source: U.S. Department of Agriculture, Miscellaneous Publication No. 572, Washington, D.C., 1945.
* Parentheses, imputed values.

Animals cannot synthesize the precursors of vitamin A, but they can convert part of some of the carotenes into vitamin A. Thus avitaminosis A can be prevented in the rat by daily administration of either 0.05 to 0.1 mg. of vitamin A or 3 to 5 mg. of carotene.

Absorption of vitamin A in the intestine is deficient when there is no bile (*e.g.*, cases of obstruction of the common bile duct); in certain intestinal infections that damage the intestinal mucosa; and when mineral oil is taken.

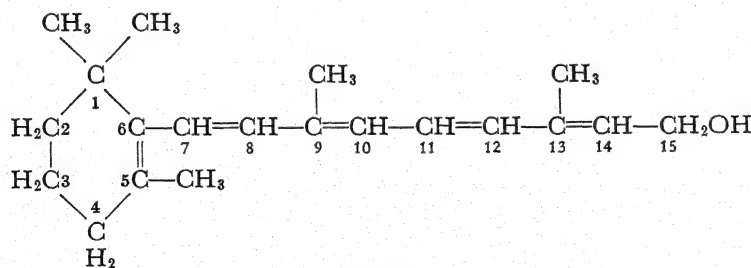
In certain cases of myxedema, carotenoids increase in the blood plasma and skin, which take on a yellow tinge. Conversion of carotenoids into vitamin A proceeds at a slow rate in these patients.

Assay. The concentration of vitamin A in foods can be determined by physical, chemical, and biological methods. The latter consist in measuring the amounts of food needed to prevent or cure avitaminosis A in the white rat. The international unit (IU), as defined in 1949, is the activity of 0.344 μg of crystallized vitamin A acetate, equivalent to 0.3 μg of pure vitamin A (alcohol). The unit of provitamin A is the activity of 0.6 μg of carotene. The two units are equivalent.

Requirement. The National Research Council recommends a daily allowance of 5,000 IU (75 IU per kg.) for an adult man or woman. Pregnant and nursing mothers and adolescents should be given a supplementary amount.

Physical and Chemical Properties

Vitamin A was first obtained in pure form by Karrer in 1931 and synthesized by Kuhn and Morris in 1937.

```
  CH3   CH3
    \   /
      C                  CH3                   CH3
    / 1 \                 |                     |
H2C2     6C—CH=CH—C=CH—CH=CH—C=CH—CH2OH
  |      ||  7   8   9 10   11  12   13 14  15
H2C3     5C—CH3
    \ 4 /
      C
      H2
```

Vitamin A

Vitamin A is soluble in fat and in organic solvents. It is easily oxidized, but it is resistant to heat and can be distilled in the absence of oxygen. It is liquid at 8°C. It shows an absorption band between 325 and 328 mμ. It gives a vivid blue color with antimony trichloride, with an absorption band between 572 and 606 mμ.

THIAMINE (VITAMIN B_1)

Thiamine, also known as the antineuritic or antiberiberi vitamin, or aneurin, is found in and produced by many yeasts, bacteria, and plants. It is an indispensable dietary factor for man and other mammals, with the exception of ruminants, and for birds. In the ruminants it is synthesized by bacteria in the digestive tract. Thiamine is essential for the growth of certain fungi and bacteria.

Avitaminosis and Hypovitaminosis

Thiamine deficiency causes many disturbances.

Nervous symptoms. There are multiple degenerative lesions in the nerves (neuritis, polyneuritis). At first there is muscular weakness, especially of the quadriceps, so that the subject has difficulty in standing up. There are also sensory disturbances, such as loss of vibratory sensation when a tuning fork is placed over the skin near a bone (tibia, malleolus, etc.). The calf muscles are sensitive to pressure and there is hyperesthesia in the feet. The ankle and knee reflexes first diminish and then disappear completely. If the deficiency is prolonged, these disturbances become permanent and the patient becomes paralyzed (Fig. 222). In acute cases there are degenerative lesions not only in peripheral nerves but also in the nerve cells and centers, *e.g.*, in pigeons there are violent convulsions and the head is thrown back (opisthotonos) or twisted; all the disorders disappear rapidly after the injection of thiamine (Fig. 223).

Thiamine plays an important part in the chemical processes of the nerve (von Muralt). It is set free when the nerve is stimulated, and it sensitizes the nerve to the effects of acetylcholine.

Cardiac symptoms. There are cardiac insufficiency with dilatation of the heart, tachycardia in man and dogs and bradycardia in rats and pigeons, and arterial hypotension with venous hypertension. The deflections of the electrocardiogram are of small amplitude and frequently the T wave is inverted (Fig. 224). There is accumulation of fluid between the cardiac fibers (edema) and vacuolization of the cardiac fibers and nerve cells in the myocardium. There are also degenerative focal lesions.[1] Intracellular and extracellular edema causes the weight of the heart to increase.

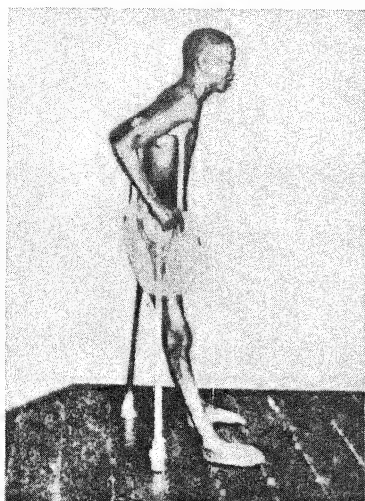



FIG. 222. Polyneuritis due to deficiency of vitamin B_1 (thiamine). Nervous form of beriberi. Muscular atrophy and impotence.

These cardiac disturbances do not respond, or respond feebly, to cardiotonic drugs if thiamine is not given simultaneously.

Digestive symptoms. Intense anorexia (loss of appetite) is usually observed. The patients do not eat enough, and inanition may occur. There is gastric achylia, and the tone of the stomach diminishes. Spasms or loss of tone are observed in the intestine.

Edema. In man there is frequently accumulation of tissue fluid in the legs, and serous effusion. The blood volume is usually decreased.

Blood and urine. The blood concentration of thiamine is low, and its excretion in the urine diminishes.

Other disturbances. Growth ceases in animals, there is considerable loss of weight owing to the lack of appetite, and the animals die emaciated after having suffered from convulsions and paralysis. Carbohydrate metabolism is disturbed; there is a characteristic increase in pyruvic and lactic acid, especially after the administration of glucose. In advanced stages, hyperglycemia and a decrease in liver glycogen may be observed, and injected glucose is not completely oxidized.

[1] SOLDATI, L. DE, "Trastornos Circulatorios de la Avitaminosis B_1," El Ateneo, Buenos Aires, 1940.

Incidence of avitaminosis B_1. Thiamine deficiency is widespread and there are millions of cases in the tropics, especially in Japan, the Philippines, India, and tropical America. The disease is called "beriberi" or "shoshin" in the East. It presents several clinical forms: (*a*) the "dry" type, with emaciation, polyneuritis, and nervous disturbances predominating; (*b*) the "wet" type, with edemas and serous effusions; (*c*) the acute type, with severe cardiac symptoms, which is rapidly fatal; (*d*) the mixed type, in which complex symptoms are observed; (*e*) the encephalopathic type.

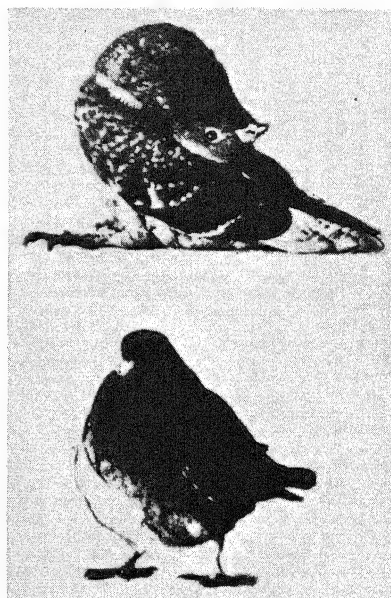



FIG. 223. Pigeons with avian polyneuritis due to deficiency of vitamin B_1 (thiamine). Opisthotonus, paralysis, and convulsions (above) disappear when thiamine is injected, and the animal has completely recovered (below) 3 hr. after the injection.

Recent experiments in man have contributed important knowledge on the initial signs and symptoms[1] of hypovitaminosis B_1 and on its

[1] WILLIAMS, R. D., *et al.*, *Arch. Int. Med.*, **66**, 785, 1940; **69**, 721, 1942; **71**, 38, 1943.

hidden forms. When only 0.10 mg. thiamine is given daily for every 1000 kg.-cal. in the diet, there are loss of appetite or nausea, weakness, tiredness, irritability, neurasthenia, precordial pain or oppression, cutaneous or muscular hyperesthesia, or hyperacousia (auditory hypersensitiveness), muscular cramps, dyspnea on exercise, insomnia, palpitations, constipation, mental torpor, loss of manual ability, and finally polyneuritis. When 0.22 mg. per 1000 kg.-cal. is allowed, in some cases there are mental and physical disturbances, but not polyneuritis. It is advisable that the diet should contain twice this amount of thiamine.

Secondary hypovitaminosis. Polyneuritis may occur in alcoholics because of deficient absorption of thiamine by the intestine. In certain gastrointestinal disorders and sometimes in pregnancy, polyneuritis is provoked by this mechanism.

SOURCES OF VITAMIN B_1

Thiamine is found in the germ and pericarp (bran) of cereals. Beriberi became prevalent in the East when the consumption of polished rice, deprived of its pericarp by milling, became general. Brewer's yeast contains large quantities of thiamine, but when it is fresh it does not always give it up easily; therefore it should be hydrolyzed or dried. Thiamine is also found in peas, beans, the leaves of green vegetables, and the following animal tissues: liver, kidney, heart, muscle, brain, and egg yolk. Milk does not contain much thiamine, and the concentration varies with the diet of the lactating female. Certain bacteria in the rumen of ruminants and the large intestine of rats and infants synthesize thiamine, which is excreted in the feces. This fact explains how coprophagia (ingestion of feces) can improve the condition or cure avitaminotic rats. Only small amounts of thiamine are stored in the body. Most of it is obtained from food or bacterial synthesis in the intestine.

FIG. 224. Cardiac dilatation due to deficiency of vitamin B_1 (thiamine). Inversion of T wave in lead IV of the electrocardiogram (left). Decrease in cardiac area and normalization of the ECG after 1 month's treatment with thiamine (right).

Assay. Thiamine can be assayed by physical, chemical, and biological methods. The most commonly used chemical method is based on the oxidation of thiamine to thiochrome; this substance gives a blue

fluorescence which can be estimated in ultraviolet light. There are several biological methods: (*a*) bacterial growth is observed in species that require thiamine for their development; (*b*) avitaminosis is provoked in rats, and the effect of the substance assayed on convulsions and paralysis or on bradycardia is observed; (*c*) the unknown substance is used for treatment of avitaminosis in pigeons.

The international unit (IU) is equivalent to the activity of 3 μg crystallized thiamine.

Requirement. The National Research Council advises an allowance of 1.5 to 2.3 mg. of thiamine daily for every adult. During the second half of pregnancy and during lactation, the maximum amount should be given. The quantity of thiamine needed varies with the diet. The optimum lies between 0.5 and 1 mg. per 1,000 kg.-cal. (not less than 0.33 mg., according to Williams). High-carbohydrate diets require more than low-carbohydrate diets of the same caloric value. The requirement increases when the metabolic rate rises, as in fever and hyperthyroidism and in pregnancy and lactation. In cases of avitaminosis B_1, 25 to 50 mg. daily should be given. These large doses are not harmful because of the low toxicity of thiamine; doses of even 500 mg. have been given without ill effects. When there is a disturbance in the intestinal absorption of vitamin B_1 (alcoholic neuritis) or it is urgent to remedy the deficiency (delirium tremens), thiamine should be injected intravenously. During the Second World War 2 mg. per lb. was added to white wheat flour, in order to compensate for possible deficiencies due to wartime dietary restrictions.

Physical and Chemical Properties

Takaki eliminated beriberi from the Japanese navy by changing the diet in 1884. A few years later Eijkman (1893–1897) observed that birds fed on polished rice suffered from polyneuritis, which could be cured by feeding them rice polishings. A crystalline substance of high antineuritic potency was obtained from rice polishings by Funk (1911). Thiamine was crystallized by Jansen and Donath in 1926, and Williams prepared it by synthesis in 1936. It is now available in large quantities at low cost.

It is formed by the union of a pyrimidine nucleus and a thiazole nucleus; it also has an amine. The name "thiamine" was suggested by its structural formula.

```
                  HCl
                   ·
         N        NH2        S
       /   \\    /          /  \
H3C—C        C          HC      C—CH2—CH2OH
     ||      |          ||      ||
     N       C          N———————C—CH3
       \   //  \       / Cl
        CH       CH2
```

Thiamine chloride

```
         N        NH2        S                          OH       OH
       /   \\    /          /  \                         |        |
H3C—C        C          HC      C—CH2CH2—O—P—O—P—O⁻
     ||      |          ||      ||                       ||       ||
     N       C          N———————C—CH3                    O        O
       \   //  \       / +
        CH       CH2
```

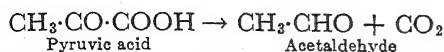
Diphosphothiamine
(Co-carboxylase)

Oxidation of α-Keto acids (pyruvic and α-ketoglutaric) in animal tissues requires a coenzyme, lipothiamide pyrophosphate, which is a combination of diphosphothiamine and lipoic acid.

Thiamine is soluble in water, in 70 per cent alcohol, and in acetic acid; it is insoluble in organic solvents. It resists dry heat but is destroyed by heat in the presence of water. It is only partially destroyed by boiling for a short time, but at 120°C. it disintegrates rapidly, especially in an alkaline solution, and more slowly in an acid medium. It is easily adsorbed by fuller's earth.

Mechanism of its Effect

Thiamine is an important factor in carbohydrate metabolism. Diphosphothiamine acts as a coenzyme of carboxylase in the decarboxylation of pyruvic acid in yeast.

$$\underset{\text{Pyruvic acid}}{CH_3{\cdot}CO{\cdot}COOH} \rightarrow \underset{\text{Acetaldehyde}}{CH_3{\cdot}CHO} + CO_2$$

When the diet is deficient in thiamine, pyruvic acid accumulates in the blood, and this accumulation inhibits the breakdown of lactic acid. Surviving brain slices taken from pigeons in avitaminosis B_1 have a low oxygen consumption, and pyruvic and lactic acids accumulate. The addition of thiamine increases oxygen consumption and formation of CO_2 and the utilization of pyruvic acid. Thiamine pyrophosphate (diphosphothiamine) acts as a coenzyme in the oxidation of pyruvic acid.

Pyruvic acid is carboxylated by animal and plant tissues, CO_2 is bound, and C_4 dicarboxylic acids are formed, which are metabolized following the tricarboxylic cycle (see Chap. 34).

Thiamine is apparently a necessary factor in the synthesis of fat from carbohydrate (McHenry). It is also a factor in water metabolism, since its deficiency causes edema: probably this effect is due to disturbances in carbohydrate metabolism.

Antivitamins. Thiaminase, an enzyme that destroys thiamine, has been found in the flesh of certain fishes. Certain substances with a chemical structure similar to thiamine (*e.g.*, pyrithiamine) interfere with he utilization of thiamine by replacing it in the enzyme systems of which it forms part. As these substances do not have the same effect as thiamine, the coenzyme is inactive and symptoms of thiamine deficiency are observed.

RIBOFLAVIN (VITAMIN B_2)

Riboflavin is synonymous with vitamin B_2 and lactoflavin.[1]

Avitaminosis and hypovitaminosis. Riboflavin deficiency in the rat causes cessation of growth, alopecia, and vascularization of the cornea. In the monkey, growth ceases and there are dermatitis and anemia. In man, ariboflavinosis is usually associated with deficiency of other vitamins in group B, especially with beriberi and pellagra; there are disturbances in the mouth, eyes, and skin.

Oral lesions. There is inflammation of the lips, which are of a bright red color and sometimes everted, especially the lower lip; frequently there is also reddening, or fissures, in the angles of the mouth (angular stomatitis). This condition is known as cheilosis (Greek χειλος, lip). There is also glossitis, *i.e.*, inflammation of the tongue, which is bright red or magenta, with flattened papillae.

Eye lesions. The eyes itch and smart. There are photophobia, pericorneal vasodilatation, and sometimes invasion of the cornea by blood vessels. In the more severe cases there is infiltration and opacity of the cornea (interstitial keratitis).

Cutaneous lesions. On an erythematous background there are scaling of the skin and accumulation of sebaceous secretion, especially in the nasolabial sulcus, in the alae nasi, around the eyelids, extending sometimes to the cheeks and forehead, and on the ears.

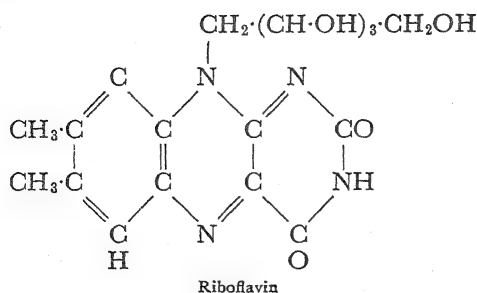
Sources and requirement. Riboflavin is one of the lyochromes, a division of the flavines, commonly found in the yellow parts of plant and animal tissues. It is also produced by certain bacteria and by brewer's yeast. It is found in milk, liver, kidney, heart, retina, and egg yolk, and in small quantities in muscles. Riboflavin exists in spinach, carrots, lettuce, peas, pears, bananas, peanuts, etc.

An adult man or woman needs from 2.2 to 3.3 mg. per day (0.6 to 0.7 mg. per 1000 kg.-cal.). The daily intake should not fall below 2 mg., and in pregnant women and adolescents it is advisable to increase it to 4 mg. In cases of ariboflavinosis, doses of 5 mg. should be administered three times per day over a long period. The elimination of riboflavin is sometimes greater than the amount ingested. This is apparently due to the production of riboflavin by intestinal bacteria.[1]

Physical and chemical properties. Kuhn *et al.* discovered the chemical nature of vitamin B_2 in 1933. It is a flavine, which can be free or bound to other substances; it has a yellow color (whence the name "flavine") and gives a green fluorescence in ultraviolet light. It has ribitol in its molecule (6-7-dimethyl-9 [*d-l'*-ribityl] isoalloxazine) and is therefore usually known as riboflavin.

[1] The term "vitamin B_2" was first used for concentrates containing several vitamins in the B group; later it was applied to riboflavin.

[1] *J. A. M. A.*, **126,** 537, 1944.

```
            CH2·(CH·OH)3·CH2OH
                 |
        C        N        N
      //  \    /   \    //  \
CH3·C       C         C        CO
    |       ||        |        |
CH3·C       C         C        NH
     \\   /   \     //  \     /
       C        N        C
       H                 O
```

Riboflavin

Combined to phosphoric acid, it forms nucleotides: (*a*) a mononucleotide, riboflavin-5′-phosphoric acid; (*b*) a dinucleotide of riboflavin and adenosine. These nucleotides act as prosthetic groups of several oxidizing enzymes which form part of the mechanism of cell respiration. They are thus of considerable importance in carbohydrate and amino-acid metabolism, oxidation of xanthine, and other chemical processes of the tissue (see Chap. 34).

NICOTINIC ACID (NIACIN) AND NICOTINIC ACID AMIDE

Certain dietary deficiencies cause a disease known as pellagra in man and black tongue in dogs. These disturbances are cured by foods that contain the so-called pellagra-preventing (P-P) factor or antipellagra vitamin or by the administration of nicotinic acid (niacin) or nicotinic acid amide (niacinamide).

COOH, N — Nicotinic acid

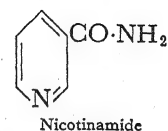
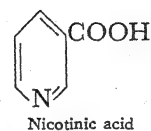
$CO \cdot NH_2$, N — Nicotinamide

Pellagra is widespread in badly nourished communities of northern Italy, Rumania, parts of the southern United States, Africa, and South America, where the main, almost exclusive,

Table 70. Content of Vitamins in Group B of a Few Typical Foods

Food, 100 *gm.*	*Thiamine, mg.*	*Riboflavin, mg.*	*Nicotinic acid, mg.*	*Pantothenic acid, mg.*	*Pyridoxine, mg.*
Apples	0.025	0.050	0.500	0.050	
Bananas	0.040	0.080	0.600	0.070	
Bread					
White unfortified	0.070	0.100	0.800	0.400	0.300
White fortified	0.280	0.140	1.500	0.400	0.300
Cabbage	0.060	0.050	0.290	0.225	0.290
Carrots	0.050	0.100	1.500	0.210	0.190
Cheese	0.030	0.500	......	0.350	
Cornmeal	0.200	0.150	1.500	0.800	
Eggs	0.250	0.400	0.050	0.700	
Meats					
Beef	0.150	0.250	6.500	1.100	0.400
Pork (loin)	1.500	0.200	9.200	1.500	0.600
Poultry (light meat)	0.075	0.060	6.100	0.800	
Poultry (dark meat)	0.100	0.250	7.300	2.000	0.200
Calf's liver	0.400	3.200	20.000	5.200	
Pork liver	0.400	2.700	22.000	5.400	
Milk (whole, fluid)	0.045	0.200	0.070	0.300	0.200
Oatmeal	0.800	0.160	1.130	1.300	0.250
Oranges	0.070	0.030	0.220	0.070	
Peas (fresh)	0.300	0.190	0.750	0.600	
Peanuts	0.800	0.300	13.000	3.400	
Potatoes	0.125	0.060	1.160	0.400	0.160
Spinach	0.075	0.250	0.720	0.200	
Tomatoes	0.050	0.050	0.580	0.075	
Turnips	0.040	0.060	......	0.250	
Yeast (brewer's, dry)	12.000	4.000	40.000	20.000	5.500
Whole wheat	0.450	0.120	5.900	1.300	0.460

Source: Elvehjem, C. A., *J. A. M. A.*, **120,** 1388, 1942.

foodstuff consumed is corn. The incidence increases in the spring and summer. Pellagra was first described by Casal in 1735 and was the object of a careful study by Goldberger. The curative effect of nicotinic acid and nicotinic amide was demonstrated in dogs with black tongue by Elvehjem in 1937 and later applied to cases of pellagra in man by Spies and Fours.

Avitaminosis and hypovitaminosis. Deficiency of this vitamin is characterized by dermatitis, diarrhea, and dementia, severe forms leading to death. The most outstanding disturbances are

1. *Glossitis.* The sides and tip of the tongue are swollen and imprinted by the teeth. The tongue is scarlet, with congested and hypertrophied papillae in the first stages and later atrophy of the papillae. Sometimes there is ulceration of the tongue. Similar lesions are observed on the lips, the gums, and the buccopharyngeal mucosa.
2. *Digestive disturbances.* There are burning sensations in the mouth, esophagus, and stomach. There are also gastric achylia, nausea, vomiting, and severe diarrhea.
3. *Dermatitis.* There is itch, and the skin reddens. Later it becomes swollen and rough, and blisters and ulcers are formed. Eventually the skin becomes thick, scaly, and pigmented. These lesions are especially marked in the face, neck, back of the hands and feet, knees, elbows, axillae, and perineum.
4. *Mental and nervous disturbances.* There are mental confusion, loss of memory, depression, and sometimes violent outbursts, with hallucinations and ideas of persecution. Many of the patients become inmates of mental hospitals. There are also peripheral nerve disturbances, and in advanced cases rigidity and other abnormalities of movement.

Sources and requirement. Pellagra in man and black tongue in dogs can be prevented or cured by the ingestion of sufficient amounts of milk, meat, and brewer's yeast. Injection of nicotinic acid or nicotinic amide is followed by rapid and sometimes spectacular results; in 24 to 72 hr. most of the disturbances completely disappear. Tryptophane has a favorable effect, as it gives rise to a certain amount of niacin, but it cannot cure the disease.

Diets that produce black tongue are deficient in several factors (protein, vitamins A, B_1, B_2, calcium), and human pellagra is usually accompanied by deficiency in thiamine (polyneuritis) and ariboflavinosis (cheilosis). Therefore in the treatment of these cases not only must nicotinic acid be given, but also other dietary deficiencies should be corrected.

Pellagra is frequently due to a deficiency of nicotinic amide, but an antivitamin has been found in corn, which competes with nicotinic amide and increases the deficiency (Wooley).

There are cases of secondary pellagra in conditions in which there is a deficient diet or disturbances in absorption. This is frequently observed in alcoholism and sometimes in malaria and tuberculosis.

A normal adult needs from 15 to 25 mg. of nicotinic acid per day (Table 74). During the Second World War, to prevent deficiencies due to dietary restrictions, 30 to 50 mg. of nicotinic acid was added to every kilogram of white flour. Patients suffering from deficiency should be given 200 to 500 mg. per day.

Mechanism of action. Nicotinic amide forms part of the pyridine-nucleotides which act as dehydrogenases in the oxidation of carbohydrate, protein, etc. (see Chap. 34). These nucleotides are (*a*) diphosphopyridine-nucleotide (DPN), also known as coenzyme I or cozymase I, formed by adenine, two molecules of phosphoric acid, two molecules of ribose, and nicotinamide; (*b*) triphosphopyridine-nucleotide (TPN), also known as coenzyme II or cozymase II, which differs from DPN in having three instead of two molecules of phosphoric acid (see Chap. 34).

FOLIC ACID

This vitamin, also known as folacin, was first called "*Lactobacillus casei* factor" or "*Streptococcus lactis* factor" because it is necessary for growth of these and other bacteria. Its deficiency, if sufficiently marked and prolonged, is fatal in several species (guinea pig); growth ceases in chicks (vitamin B_c), and paralysis develops in turkeys. Anemia is provoked by folic-acid deficiency in several animals, and certain forms of anemia in the chick, rat, and monkey and in man (pellagra, sprue, pregnancy, etc.) are improved by folic acid. Some cases refractory to treatment with vitamin B_{12} respond to folic acid. In some cases of pernicious and other macrocytic anemias in man, folic-acid treatment causes the anemia to disappear, but nervous

disturbances due to degenerative lesions in the spinal cord are not modified (see Chap. 5).

Folic acid is found in the liver and kidneys, in eggs and milk, and in many seeds and leaves (hence its name).

Folic acid has been prepared by synthesis. It is also known as pteroylglutamic acid, and its component parts are pteridin, paraminobenzoic acid, and glutamic acid. It is similar to the yellow pigment in the wings of butterflies.

Folic acid is converted by the liver into folinic acid or citrovorum factor, or leucovorin, so called because it is necessary for growth of *Leuconostoc citrovorum*. This factor forms part of a coenzyme essential to formate (HCOOH) transfer. This process is an important step in the synthesis of purines and thymine, and therefore in that of nucleic acids. Formate carbon also appears in serine, choline, and methionine.

Certain substances chemically related to pteroylglutamic acid, *e.g.*, 4-aminopteroylglutamic acid (aminopterin) act as folic-acid antagonists, *i.e.*, they provoke symptoms of folic-acid deficiency which are reversed by large doses of folic acid, and more efficiently by folinic acid.

VITAMIN B_{12}

This vitamin is found in several foods, but it is not absorbed in the absence of Castle's intrinsic gastric factor. It is Castle's extrinsic factor (see Chap. 5). It was first obtained in 1948 from liver extracts, but is now prepared on a large scale from cultures of a mold, *Streptomyces griseus*. It forms red crystals which contain cobalt. Three of these vitamins have been separated: B_{12}a, which contains a cyanide group (cyanocobalamine); B_{12}b, which does not have this group (hydroxycobalamine); and B_{12}c (nitrocobalamine).

It not only has a potent antianemic activity (see Chap. 5) but also is a growth factor for the chick, rat, pig, etc., and for several bacteria, *e.g.*, *Lactobacillus lactis*. It has antithyroid and lipotropic effects. Vitamin B_{12} plays a part in protein anabolism, in the synthesis of purines and thymidine, and in transmethylation. Its activity is in some way related to that of folic acid.

OTHER VITAMINS IN GROUP B

At first it was thought that yeast contained a single vitamin, which was called "vitamin B," but gradually several different factors necessary for growth and normal metabolism in different animal species, yeasts, and bacteria were separated from the "B complex." The factor necessary for the growth of yeast was called "bios," but later this was found to be made up of inositol and biotin. The factors essential for bacterial growth can be inhibited by several other substances known as antibiotics (such as the sulfanilamides), which thus suppress the development and pathogenic effects of many bacteria. The following is a short list of vitamins that are active in different animals.

Vitamins B_3 to B_8. These are vitamins necessary for normal nutrition in different species. B_3 is mostly pantothenic acid, essential in pigeons. B_4 is necessary to rats and B_5 to pigeons. B_8 is adenylic acid.

Pyridoxine. Vitamin B_6 has been crystallized and identified as 2-methyl-3-hydroxy-4,5-dihydroxymethylpyridine. The aldehyde (pyridoxal) and amine (pyridoxamine) are also active. This vitamin is a growth factor for rats and for many bacteria. Pyridoxine deficiency in rats causes acrodynia (painful extremities) and dermatitis, which are cured by the administration of pyridoxine or one of its derivatives. In dogs and pigs there is also anemia. At one time it was mistakenly believed that vitamin B_6 was the antipellagra factor for man. It does not cure pellagra or beriberi, but improves the condition of the patients by rapidly diminishing the severity of certain symptoms, such as irritability, insomnia, abdominal pains, disturbances in movements, and ulcerations resistant to nicotinic amide, thiamine, and riboflavin. Pyridoxine has been found in human milk and urine, but deficiency is not easily provoked in man.[1] It is thought an adult needs 1.5 to 2 mg. daily.[2] Pyridoxal phosphate acts as a coenzyme of some amino-acid decarboxylases and transaminases, and is indispensable for the metabolism of tryptophan.

Biotin. Du Vigneaud crystallized and synthesized biotin in 1940.[3] It is necessary for the growth of certain bacteria, and it acts in several carboxylation and decarboxylation reactions. Rats fed on egg white have dermatitis and eventually die, unless they are given large doses of biotin, because a substance known as avidin, found in egg white, combines with biotin and suppresses its activity. In the dog and rat, biotin deficiency causes paralysis and death. Biotin has an antilipotropic effect; it produces fatty infiltration of the liver. It has been found in human

[1] Mueller, J. E., and R. W. Vilter, *J. Clin. Investigation*, **29**, 139, 1950.
[2] Williams, R. J., *J. A. M. A.*, **119**, 1, 1942.
[3] Du Vigneaud, V., *Science*, **96**, 454, 1942.

tissues and in blood, but its importance in human nutrition is not yet well understood. Biotin deficiency in man[1] causes paleness of the skin and mucosae (due to slight anemia), cutaneous desquamation, loss of appetite, tiredness, and muscular pains. An adult requires 0.15 mg. of biotin daily.

Inositol. Deficiency of this substance causes retardation in growth and loss of hair in rats. Sometimes only the hair around the eyes is lost, and the animals appear to be wearing spectacles. It is a lipotropic factor and prevents or cures the fatty liver provoked by biotin. Its importance in human nutrition is unknown; nevertheless Williams believes the diet of an adult should contain 1 gm. per day.

Para-aminobenzoic acid (PABA). This substance is necessary for the growth of several bacteria, probably because it is needed for the formation of folic acid and related substances. Sulfanilamides apparently compete with PABA, thus checking bacterial development. It has a preventive effect on certain rickettsias. PABA deficiency causes graying of the coat in black rats (achromotrichia), and the normal color is restored by treatment with PABA, but it does not restore it in human graying hair, as was believed at one time.

Pantothenic acid. This substance is found in nearly all animal tissues; hence its name. It has been obtained in pure form and was synthesized by Williams and his associates in 1940. It is a growth factor in certain bacteria, yeasts, and plants. Together with adenosine, thioethanol amine, and three phosphate groups it forms coenzyme A which acts in transacetylation processes in the formation of citric acid, acetylcholine, etc. Its deficiency causes dermatitis and loss of color in the plumage of chicks, and graying of the coat in mice. In the rat there are retardation of growth, anemia, necrosis of the adrenals and other organs, and several other disturbances. Pantothenic acid has been found in the blood of man; it diminishes in thiamine and riboflavin deficiencies and in pellagra. Its importance in human nutrition is not yet well known, but Williams advises that 11 mg. per day for an adult man or woman should be allowed in a normal diet.

VITAMIN C (ASCORBIC ACID)

Vitamin C is also known as ascorbic acid or cevitamic acid; formerly it was called the antiscorbutic vitamin. It was found in the adrenal cortex by Szent-Györgyi (1928), who demonstrated its antiscorbutic activity in 1932. King and Waugh obtained it from lemon juice.

Avitaminosis and hypovitaminosis. Scurvy is due to ascorbic-acid deficiency; man and the guinea pig are the most sensitive species. Experimental scurvy can be easily provoked in guinea pigs fed on a scorbutic diet. After 3 weeks there are loss of weight, subperiosteal hemorrhages, lesions in bones and teeth, and digestive disturbances; the teeth can be extracted almost without effort. Eventually the animals die. In man scurvy has been observed in small children fed exclusively on milk sterilized by heating, and in adults fed for a long time on a restricted diet without fresh food. Scurvy was observed in sailors on long voyages in the time of sailing ships, in members of polar expeditions, and in the inhabitants of besieged towns. The most recent occurrences of scurvy on a wide scale were observed in the besieged city of Kut-el-Amara in Mesopotamia, during the First World War, and in the war in the Chaco between Paraguay and Bolivia. In the eighteenth century the English naval surgeon Jones demonstrated that lemon juice and fresh vegetables prevent or cure scurvy. The severity and duration of scurvy depend on the degree of the deficiency of ascorbic acid. In man the following are the main signs and symptoms of the disease:

Hemorrhage frequently occurs in the skin, under the periosteum, in the mucosae of the mouth, in the digestive tract, in the serous cavities including the joint spaces (painful joints), and in the muscles.

The gums are red, swollen, sensitive, and bleed easily. In chronic cases of scurvy they are thickened and retracted, leaving bare the roots of the teeth or forming a cleft between the tooth and the gum which, on becoming infected, gives rise to pyorrhea.

The teeth soon suffer severe damage. Periodontal tissues are softened and teeth fall out spontaneously (expulsive gingivitis) or when the slightest force is exerted on them. Dentine is porous and poorly calcified. The pulp is congested; the odontoblasts are irregularly distributed and show degenerative lesions. In chronic scurvy, osteodentine is formed in the pulp cavity.

The bones show varied lesions. In children the lower limbs are painful, there are subperiosteal hemorrhages, and endochondral ossification

[1] SYDENSTRICKER, V. P., *et al.*, *J. A. M. A.*, **118**, 1199, 1942.

ceases, because the organic matrix is not formed. After a time there is osteoporosis, and in chronic cases there are marked lesions in the bone and in the neighborhood of the joints. Fractures are not easily consolidated.

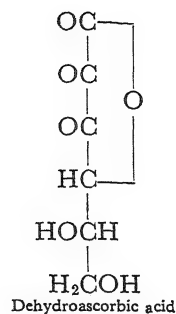
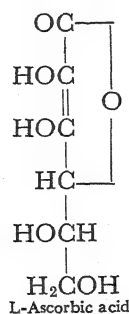
Anemia, general weakness, and digestive disturbances are observed in scurvy. Diarrhea and sometimes pneumonia precede death in severe cases.

The concentration of ascorbic acid in blood falls from the normal of 1.2 mg. per cent to 0.5 mg. per cent or less. Urinary excretion of ascorbic acid is also considerably below the normal of 30 to 50 mg. daily. If 300 mg. of ascorbic acid is given to a normal subject, 50 mg. or more is excreted in the urine in the course of the next day. Patients suffering from scurvy retain ascorbic acid, and the dose must be repeated for several days (saturation period) before urinary elimination of ascorbic acid takes place as in normal subjects. The examination of blood ascorbic acid, or its elimination, and testing of capillary resistance, are the methods usually employed to detect slight degrees of ascorbic-acid deficiency.

Sources and requirement. Ascorbic acid is found in large quantities in (*a*) citrus fruits, such as lemons, limes, oranges, and grapefruits (50 to 100 mg. per 100 gm.), and strawberries (50 to 70 mg. per 100 gm.); (*b*) fresh vegetables, such as cabbages (30 to 120 mg. per cent), green and red peppers (100 to 200 mg. per cent), tomatoes (10 to 40 mg. per cent), spinach (8 to 33 mg. per cent), fresh peas (40 to 50 mg. per cent), potatoes (8 to 25 mg. per cent); (*c*) seeds, after the onset of germination (not before). Cow's milk has little vitamin C; there is more in human milk. In boiled milk most of it is destroyed.

The optimum amount for an adult is 70 to 100 mg. per day (Table 74). It should never fall below 10 mg. per day, and it is better to allow more than 30 mg. for adults. Small children should be given 3 to 8 mg. per kg. of body weight per day, adolescents 6 to 7.5 mg. per kg., and adults 0.7 to 1.6 mg. per kg. In cases of scurvy, doses between 250 mg. and 1,000 mg. per day have been given without observing toxic symptoms.

Physical and chemical properties. In animal tissues L-ascorbic acid is found in equilibrium with dehydroascorbic acid. It has six carbon atoms, and its structural formula is as follows:

```
  OC——┐             OC——┐
  |   |             |   |
 HOC  |             OC  |
  ||  O             |   O
 HOC  |             OC  |
  |   |             |   |
  HC——┘             HC——┘
  |                 |
 HOCH              HOCH
  |                 |
 H2COH             H2COH
L-Ascorbic acid    Dehydroascorbic acid
```

Ascorbic acid is destroyed by prolonged boiling, owing to progressive oxidation. Copper catalyzes this oxidation. Drying destroys most of the ascorbic acid in foods, unless it is performed by special rapid processes that prevent oxidation. Ascorbic acid is preserved better in citrus fruits than in other natural foods.

Assay. Biological assay of ascorbic acid is performed by measuring the preventive or curative effect of the substance tested on experimental scurvy in the guinea pig, or by histological demonstration of the protective effect on the teeth (Höjer). Chemical methods are based on the reducing power of ascorbic acid, measured by the decoloration of methylene blue (Martini and Bonsignore), or more usually by Tillman's method, which measures the decoloration of 2-6-dichlorophenolindophenol. The latter method gives erroneously high results for leaves that contain tannin (tea, maté, etc.).

The international unit is the activity of 0.05 mg. L-ascorbic acid.

Mechanism of action. Ascorbic acid acts as a reducing agent *in vitro*. This fact led to the belief that it played a part in oxidation-reduction processes in the body, as a hydrogen carrier, but there is not yet sufficient proof of this. On the other hand there is satisfactory evidence that in ascorbic-acid deficiency intercellular substances are not normally formed. Collagen fibrils and the organic matrix of bone are not formed, and the capillary endothelium becomes fragile. Ascorbic acid exerts some protective effect against intoxication by lead, arsenic, benzene compounds, and certain bacterial toxins.

Antivitamin. Glucoascorbic acid inhibits the effect of ascorbic acid and produces scurvy not only in guinea pigs but also in the rat, an animal that does not require ascorbic acid in the diet because it can produce it by synthesis in its body.

Vitamin P and rutin. Vitamin P, or citrin, is found together with ascorbic acid in the juice of citrus fruits and other vegetable foods. Deficiency of this vitamin, according to Szent-Györgyi, is in part responsible for hemorrhages observed in scurvy, because it controls capillary permeability. The physiologic importance of this vitamin is still the subject of discussion. Substances with vitamin P activity are flavone derivates; the most active one used in therapeutics is rutin, obtained from rue (*Ruta graveolens*). Treatment with these substances diminishes capillary fragility.

FIG. 225. Radiograph of the wrist of a 14-month-old child suffering from rickets (left). After treatment with vitamin D the radiograph (right) showed normal calcification. (*Courtesy of Professor Garrahan.*)

THE D VITAMINS

The D vitamins are sterols which regulate certain processes of calcium and phosphorus metabolism, especially the formation of bone and teeth, and which prevent or cure rickets. There is not only one vitamin D; up to the present, 10 provitamins D are known, and four vitamins D have been obtained in pure form. The provitamin most frequently found in higher animals and man is 7-dehydrocholesterol, which is converted into vitamin D_3 by irradiation with ultraviolet rays. The predominant provitamin in yeasts and lower organisms is ergosterol, which is converted by ultraviolet irradiation into vitamin D_2 and is generally called calciferol, viosterol, or vigantol.[1] It is the one most frequently used in the treatment of rickets.

AVITAMINOSIS AND HYPOVITAMINOSIS

Deficiency of vitamins D causes rickets. Children from four months to two years of age are more susceptible than older children. It is a disease that involves the whole body; the most outstanding signs and symptoms are seen in the bones and teeth and in phosphorus metabolism. It is produced only during the period of growth. Slight forms do not disturb the general health of the child; the common idea of a rickety child, emaciated and with grossly deformed bones, is true only of severe rickets.

Bone lesions. There are marked disturbances in endochondral ossification, but the whole bone is involved. The epiphyseal cartilages are enlarged, especially those of the wrists, ankles, knee, elbow, and chondrocostal joints ("beading of the ribs" or "rachitic rosary") (Fig. 225). The fontanels remain open up to an advanced age. The bones of the head are soft and can be easily depressed (craniotabes); bosses are developed in

[1] Vitamin D_1 is not a chemical entity; it is a combination of vitamin D_2 and lumisterol. Vitamin D_4 is obtained by irradiation of 22-dehydroergosterol. Vitamin D_5 is prepared by irradiating 7-dehydrositosterol.

the frontal and parietal bones. The bones are poorly calcified, soft, and pliable. The natural curvatures of the long bones are exaggerated (knock-knees, bowlegs, curvature of the spine, etc.). In severe rickets, the thorax and pelvis are deformed.

Microscopic examination shows a disturbed and retarded endochondral and periosteal ossification. The epiphyseal cartilage grows to ten times or more its normal width. The cartilage cells are irregularly distributed, not in orderly columns as in normal bone. Osteoid, poorly calcified tissue is formed, and the bony trabeculae are weak and irregular. In severe rickets, calcium and phosphorus are reabsorbed from the bone, which is thus softened and becomes deformed by the stress and strain exerted by the body weight and muscles.

Dental lesions. In severe rickets, dentition is delayed, and there is defective implantation and malformation of the teeth. Calcification of enamel and dentine is deficient. The enamel surface is not smooth and bright, but irregular and pigmented, and in places the enamel is missing (hypoplasia). The pulp cavity is enlarged owing to poor development of dentine, which is not compact and homogeneous; it is incompletely calcified. Calcium is deposited in the organic matrix forming spheres (calcospherites), which do not coalesce as in normal teeth but leave spaces filled with organic substance between them (globular dentine) (Fig. 226).

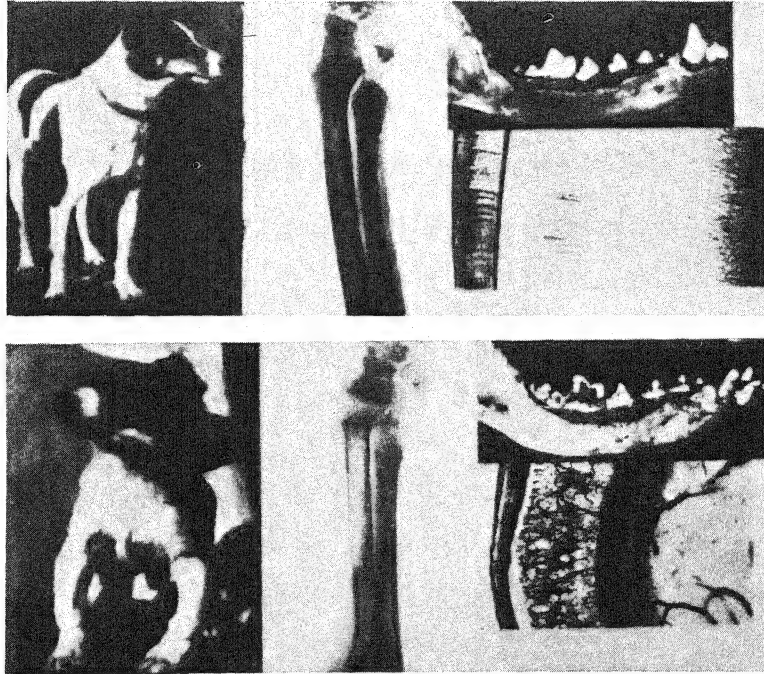



Fig. 226. Bone and dental lesions in rickets. Above: normal dog, bones, jaw, and section of tooth showing homogeneous dentine. Below: dog with experimental rickets, bowlegs, poorly calcified bones, hypoplasia in teeth, globular dentine, and defective enamel. (*Mellanby, M., "Diet and the Teeth," Med. Res. Council. Spec. Report No. 191, London, 1934.*)

General symptoms. Muscular weakness and hypotonia are observed in cases of severe rickets. Frequently there is a certain degree of anemia.

Metabolic disturbances. There are significant changes in phosphorus metabolism. Inorganic phosphate in plasma falls from between 4 and 6 mg. per cent (the normal level in children) to 2 mg. per cent or less. Phosphorus is mobilized from the bones and its excretion is abnormally high.

Plasma calcium is usually normal (9.5 to 11 mg. per cent). In some cases it decreases slightly; if it falls below 7 mg. per cent tetany may be produced. Chemical analysis of rachitic bone shows that it contains less Ca and P than normal bone, and has a higher percentage of water and organic substances.

Incidence. Rickets occurs frequently in communities deficiently nourished and not much exposed to sunlight; it is especially prevalent in children kept indoors in dark rooms. Rickets is increased by the inclusion in the diet of substances such as beryllium, iron, aluminum, strontium, or an excess of calcium, which insolubilize phosphorus and decrease its absorption, or substances that insolubilize calcium, such as phytin and cereals that contain phytin.

Mild rickets is frequently ignored. It can be detected only by taking x-ray pictures of the wrist and by the determination of phosphate and phosphatase in plasma. A complete study of experimental rickets requires also a study of the calcium and phosphorus balance, chemical analysis of the bone, and histologic examination of the epiphyseal cartilages.

Sources and Requirement

The D vitamins are sterols found in the unsaponifiable fraction of many fats. There are considerable amounts in several fish-liver oils, in butterfat, in egg yolk, and in irradiated yeast. The most commonly used source is cod-liver oil, but other fish-liver oils have 100 to 400 times the concentration of vitamin D of cod-liver oil. The D vitamins in fish-liver oils are mixtures of vitamins D_2, D_3, and others less well known. Milk, cream, and butter are also sources of these vitamins. Egg yolk contains vitamin D, and the concentration increases if the eggs have been laid by hens fed with ergosterol or irradiated with ultraviolet rays.

Ultraviolet rays activate provitamins in the body or in foods and convert them into one of the D vitamins. Rachitic children have been treated successfully with ultraviolet irradiation (Huldschinsky), which converts a provitamin in the skin into vitamin D_3. Ultraviolet irradiation of milk, yeast, and other foodstuffs has been used to increase their antirachitic effect by converting the provitamins they contain into the corresponding D vitamin.

Sunlight is of great importance in the prevention of rickets, but it is much less efficacious than ultraviolet rays in the treatment of this disease.

The amount of vitamin D in a substance can be determined by spectrophotometric analysis, because D vitamins have characteristic absorption spectra. Biological methods consist in estimating the curative effect on rats with rickets provoked by feeding a diet lacking vitamin D. In this case the effect should be compared with that of a standard preparation, because of variations in the response of the animals.

The international unit is the activity of 0.025 μg of crystallized calciferol (1 gm. is the equivalent of 40,000,000 units).

Requirement. Children, women in the second half of pregnancy, and nursing mothers should be given 800 IU daily, and the amount should never fall below 400 IU. The requirement of adults is not well known. Artificially fed infants should always be given cod-liver oil or some other vitamin D preparation, and their milk should be reinforced by the addition of 400 IU irradiated ergosterol per quart or by irradiation with ultraviolet rays. Vitamin D has sometimes been added to flour (not less than 250 nor more than 1,000 IU per kilogram of flour).

Treatment of rickets. The administration of vitamin D in patients suffering from rickets diminishes the excretion of calcium and phosphorus, and increases blood phosphate. It causes the deposition of Ca and P in bone, and normal ossification is restored. In some cases with low blood calcium, the sudden increase in blood phosphorus may provoke tetany because the P: Ca ratio is altered.

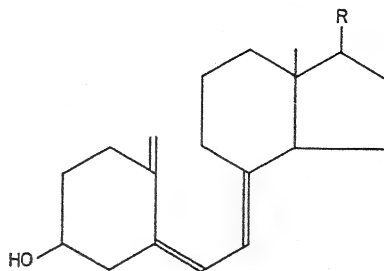
Physical and Chemical Properties

The D vitamins are sterols with a cyclopentene-phenanthrene nucleus; the B ring is opened by ultraviolet irradiation. Vitamin D_2 or calciferol is derived from ergosterol, and vitamin D_3 is obtained by irradiation of 7-dehydrocholesterol.

Commercial preparations of vitamin D are usually solutions of calciferol. Irradiation of ergosterol with ultraviolet rays provokes the formation of a series of substances in the following order: ergosterol → lumisterol → tachysterol → calciferol (vitamin D_2) → toxiferol and other derivatives.[1] The D vitamins crystallize; they are colorless, are soluble in fats, and can be distilled in a vacuum. They have typical ultraviolet-light absorption spectra.

[1] Dihydrotachysterol (commercially known as AT 10) has little antirachitic activity, but it is useful in the treatment of tetany because it raises the blood calcium, an effect which is obtained by treatment with vitamin D only when very large doses are given (150,000 to 400,000 IU per day, according to Severinghaus).

Mechanism of action. Vitamins D have the following effects: (*a*) they increase Ca and P absorption in the intestine; (*b*) they control deposition of P and Ca in the bone, and so regulate normal ossification; (*c*) they decrease elimination of Ca and P by the intestine, especially in patients with rickets; (*d*) they control the blood phosphate level. Their action differs considerably from that of parathyroid hormone (see Chap. 55) which raises blood Ca, lowers blood phosphate, and mobilizes Ca in the bone, causing an increase in calcium excretion, a negative calcium balance, and decalcification of bone. Vitamin D in moderate (physiologic) amounts, on the contrary, increases absorption and retention of calcium; only very large doses cause a rise in blood calcium. Parathyroidectomy does not suppress the effects of vitamin D. On the other hand the parathyroids are hypertrophied in rickets.

Common formula of provitamins D

Common formula of vitamins D

Vitamin D_3

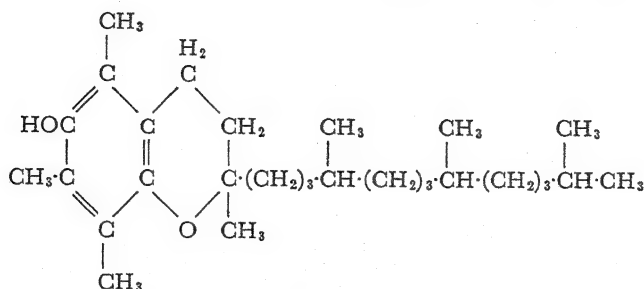
Toxicity. Pure vitamin D_2 has little toxicity. Cases of intoxication observed a few years ago were due to toxic impurities in vitamin D preparations. More recently toxic symptoms have been observed only after prolonged treatment with high doses (20,000 IU per day). Single doses of 600,000 and even 1,000,000 IU have been given to patients with rickets with excellent therapeutic effects and without provoking toxic symptoms. In tetany due to parathyroidectomy, an initial dose of 10 to 40 mg. (400,000 to 1,600,000 IU) and subsequent daily doses of 3 to 5 mg. (120,000 to 200,000 IU) have been administered without producing toxic effects, hypercalcemia, or renal disturbances. These doses maintain a normal blood-calcium level and cure tetany.

Experimental or clinical vitamin D intoxication, owing to prolonged administration of high doses, is accompanied by the following signs and symptoms: (*a*) loss of appetite, vomiting, and diarrhea; (*b*) hypercalcemia, in which plasma calcium may reach a level of 17 mg. per cent; (*c*) in the first stages, retention of Ca, which goes on to decalcification and excess elimination of Ca and P; (*d*) abnormal deposition of calcium in the kidney, stomach, arteries, etc. Severe cases have ended in death.

THE E VITAMINS

Vitamin E has also been called the antisterility vitamin. It is indispensable in several animal species, but there is no evidence that it is needed by man, and so far no human cases of vitamin E deficiency

have been reported. Evans and Bishop discovered it in 1922, and Evans and the Emersons obtained it in pure form in 1938. It has been prepared synthetically by Karrer and his associates (1938). Three substances are known to have vitamin E activity: α-tocopherol, β-tocopherol, and γ-tocopherol. The first of these is the most active of the three.

```
          CH3
           |      H2
           C      C
         //  \  /   \
     HOC       C      CH2      CH3        CH3        CH3
       |       ||     |        |          |          |
   CH3·C       C      C·(CH2)3·CH·(CH2)3·CH·(CH2)3·CH·CH3
         \\  /   \  / |
           C      O   CH3
           |
          CH3
```

α-Tocopherol

Vitamin E deficiency causes sterility in the rat. In the male there is testicular atrophy with irreversible degeneration of the germinal epithelium. In the female the sexual cycle, ovulation, copulation, and fertilization are normal, but between the eighth and twentieth day of pregnancy the blood vessels of the placenta are occluded by proliferation of mesodermal cells, causing death and reabsorption of the fetuses. In prolonged deficiencies paralysis has been observed in young rats, muscular dystrophia in several species, and softening of the brain in chicks. All these disturbances are prevented or cured by the administration of vitamin E, which also prevents liver necrosis provoked by dietary deficiencies. Tocopherols act as antioxygens in the oxidation of body fats, and they protect vitamin A from oxidation. In man, treatment with vitamin E has been given in cases of repeated abortion and in muscular dystrophia. Remarkable results have been reported, but they have not been confirmed; more evidence is needed before arriving at definite conclusions. Vitamin E is found in wheat-germ oil, egg yolk, alfalfa, lettuce, and other foodstuffs. The international unit is the activity of 1 mg. of synthetic α-tocopherol acetate in oil.

VITAMIN K

Vitamin K deficiency causes a tendency to hemorrhage, which is due to a decrease in the formation of prothrombin. Dicumarol, salicylate, and some of the naphthoquinones prepared by synthesis inhibit the effect of vitamin K. This vitamin was considered in Chap. 8, when dealing with the coagulation of the blood.

CHOLINE

Choline is not strictly a vitamin, but its importance in nutrition is now well understood.[1] Deficiency in choline causes a series of disturbances: (*a*) fatty infiltration of the liver, which is followed by cirrhosis if the condition lasts for some time; (*b*) degenerative lesions and hemorrhagic necrosis in the cortex of the kidney (if the animal is young and recovers, later hypertension may develop); (*c*) delayed growth; (*d*) papillomatous epithelial hyperplasia of the stomach in the rat; (*e*) atrophy of the thymus; (*f*) diminished egg laying in hens; (*g*) in chicks and turkeys, a disturbance known as the slipped-tendon syndrome, or perosis; (*h*) diminished milk secretion. If the deficiency is sufficiently prolonged, the animals die.

All these disturbances are prevented or cured by the administration of choline or other substances which are methyl donors, *i.e.*, easily release methyl groups, such as betaine, methionine, etc. (see Chap. 42, Fat Metabolism). These substances have lipotrophic activity; they prevent or cure fatty infiltration of the liver. Choline increases the production of phosphatides in the liver and enters into the constitution of lecithin. Moreover, choline is the parent substance of acetylcholine, which is of great importance in the transmission of the nerve impulse (see Chap. 84). Several substances give up methyl groups to each other in a process known as transmethylation (see "Transmethylation," Chap. 42).

Fatty infiltration of the liver and cirrhosis of the liver have been treated with some success

[1] Best, C. H., and C. C. Lucas, Choline, *Vitamins & Hormones*, 1, 1, 1943.

by the administration of protein, choline, or methionine.

BIBLIOGRAPHY

American Medical Association, "Handbook on Nutrition," Chicago, 1951.

Annual Review of Biochemistry, Stanford University Press, Stanford, Cal.

Aschkenasy, A., La Vitamine B_{12} et le facteur des proteines animales, *Ann. de la nutrit. et de l'aliment.*, **4**, 141, 1950.

Association of Vitamin Chemists, "Methods of Vitamin Assay," Interscience Publishers, Inc., New York, 1942.

Coward, K. H., "Biological Standardization of the Vitamins," Baillière, London, 1947.

Eddy, W. H., "Vitaminology, The Chemistry and Function of the Vitamins," Baillière, London, 1950.

Ergebnisse der Vitamin- und Hormonforschung.

György, P., "Vitamin Methods," Academic Press, New York, 1950–1951.

Harris, L. J., "Vitamins, A Digest of Current Knowledge," J. & A. Churchill, London, 1951.

Horwitt, M. K., E. Liebert, O. Kriesler, and P. Wittman, "Investigations of Human Requirements of B-complex Vitamins," National Research Council, National Academy of Science, Washington, 1940.

Jukes, T. H., and E. L. R. Stokstad, The Role of Vitamin B_{12} in Metabolic Processes, *Vitamins & Hormones*, **9**, 1, 1951.

Robinson, F. A., The Vitamin B Complex, Wiley, New York, 1951.

Rosenberg, H. R., "Chemistry and Physiology of the Vitamins," Interscience Publishers, Inc., New York, 1942.

Spies, T. D., and H. B. Butt, Vitamins and Avitaminoses, in Duncan, "Diseases of Metabolism," Saunders, Philadelphia, 1947.

Villela, G. G., "Vitaminas, Métodos de Dosificación," El Ateneo, Buenos Aires, 1948.

Vitamins and Hormones, Academic Press, New York.

CHAPTER 50

The Normal Diet of Man

THE FEEDING OF human beings has improved considerably in the last few decades, but it is still far from satisfactory in many parts of the world. An adequate diet is of primary importance in the maintenance of the health and efficiency of a people. If the diet is not up to rational nutritional standards, there is an unfavorable effect on growth, vigor, and health; there is high infantile mortality, resistance to disease is lowered at all ages, and the life span is shortened.

Even in prosperous communities part of the population is inadequately fed, and in less prosperous areas there is frequently widespread and even serious malnutrition. The principles of nutrition are universally valid, but limitation of resources in some cases, and habit and taste everywhere, are major factors in the establishment of the diet in different countries.

The problems of nutrition are often extremely complicated because they involve a knowledge not only of chemistry, biology, medicine, and public health, but also of international economics and finance, agriculture, industry, commerce, and transport. They frequently require the aid of the psychologist, not to mention the basic importance of the culinary art. Bad cooking spoils not only the taste of food but also its nutritive value. This is not the place to discuss all these problems; here only the requirements of man in calories, protein, fat, carbohydrate, mineral substances, and vitamins in different circumstances will be considered. The diet should contain a sufficient amount of the substances needed to fulfill the requirement of energy and to permit optimum growth, maintenance of body temperature, regulation of nutrition, and good health.

Food supplies the organism with (*a*) energy for the fulfillment of normal functions; (*b*) substances for growth, maintenance, and repair of body tissues; (*c*) regulatory substances, which in minimal amounts are indispensable for normal nutrition. Food therefore has three functions: energetic, plastic, and regulatory. In practice the most frequently encountered deficiencies are those of vitamins, mineral substances, and protein (indispensable amino acids).

Foodstuffs have been divided into protective and energetic. Protective foods are those which contain relatively large amounts of vitamins, minerals, or proteins that are otherwise scarce or absent in the food most commonly consumed. Thus in the United States the diet consists in great part of white bread, meat, and sugar; protective foods, therefore, are milk and green vegetables. In certain districts of Asia the main constituents of the diet are rice, soya beans, and a few green vegetables; meat has great protective value. The most important protective foods are milk and milk products (butter, cream, cheese, etc.), eggs, glandular tissues (liver, kidney, etc.), green vegetables and fruit, fish, and meat (muscle).

Energy-providing foods yield a large quantity of calories, but they are usually deficient in vitamins, mineral substances, or proteins. Thus cereals provide calories and have some protective action, but this protective action is lost by milling, and the more highly milled flours have no protective effect. Sugar and starch provide only energy; they contain neither minerals nor vitamins. If nonprotective foods form a large part of the diet, the amount of protective foods may not be enough to provide the indispensable amino acids, vitamins, and minerals. Potatoes are among the best of energy-providing foods, because they contain protein, vitamins C and B, calcium, phosphorus, and iron.

THE DIETARY REQUIREMENT OF MAN

The human diet should fulfill the following conditions:

1. It should be digestible, agreeable, and well prepared. A diet that is excellent for a herbivorous animal may not be digested by man. The aspect and taste of food should awaken appetite, thus stimulating secretion and motor activity in the digestive tract. Good cooking not only increases the agreeableness of food but also facilitates digestive processes by softening or partial hydrolysis; it also destroys some parasites that may have contaminated the food. Cooking must not destroy vitamins or wash out mineral substances from the food.
2. It should provide a sufficient amount of energy to cover the caloric needs of the organism.
3. It should contain protein, fat, and carbohydrate; in particular, a sufficient quantity of the indispensable amino acids.
4. It should contain minerals in sufficient amounts, particularly calcium, phosphorus, iron, iodine, chloride, sodium, and potassium.
5. Vitamins should be included in sufficient quantities.
6. There should be an adequate balance between the different kinds of food to assure optimum nutrition.
7. The final object of nutrition is to fulfill all the needs of the organism so it may function harmoniously as an integrated unit.

Special needs created by climatic, geographic, or individual conditions should be taken into account. Age, sex, height, weight, pregnancy, lactation, nature of work performed, environmental temperature, and level of the BMR are some of the factors that modify dietary requirements. Moreover, a diet must be balanced; the relative need of some foodstuffs varies with the quality and quantity of other components.

The diet must contain a minimum amount of each one of the indispensable principles; otherwise disturbances are observed. This amount can be increased to an optimum, but an excess may cause serious damage to the organism. For example, water is one of the simplest and most indispensable of foods, and man cannot live for long if totally deprived of it, yet excessive amounts of water will produce the syndrome of

Table 71. Composition and Caloric Values of Selected Foods

Food, 100 *gm.*	*Protein, gm.*	*Fat, gm.*	*Carbohydrate, gm.*	*Calories*
Meat, poultry, fish				
Beef				
Roasting meat, medium	18.9	13.0	0	193
Very fat choice prime, carcass, side, incl. kidney fat	13.7	39.0	0	406
Loin steaks	16.9	25.0	0	293
Lamb				
Carcass, side; thin	17.1	14.8	0	202
Carcass, side; fat	13.0	39.8	0	410
Leg roast	18.0	17.5	0	230
Pork				
Packer's carcass; side; thin	14.1	35.0	0	371
Packer's carcass; side; fat	9.8	55.0	0	534
Ham (fresh)	15.2	31.0	0	340
Poultry				
Chicken, boned, canned	21.8	9.8	0	175
Chicken, roasters	20.2	12.6	0	194
Turkey, medium fat	20.1	20.2	0	262
Fish				
Cod	16.5	0.4	0	70
Salmon, canned	20.6	9.6	0	169
Sardines, canned in oil, drained solids	25.7	11.0	1.2	207
Sardines, canned in oil, total content of can	21.1	27.0	1.0	331
Milk and milk products				
Milk				
Fresh whole	3.5	3.9	4.9	69
Fresh skim	3.5	0.1	5.1	35
Dry whole	25.8	26.7	38.0	496
Dry skim	35.6	1.0	52.0	359
Cream (20%)	2.9	20.0	4.0	208
Cheese				
Cheddar type	23.9	32.3	1.7	393
Cream	7.1	36.9	1.7	367
Fats, oils				
Butter	0.6	81	0.4	733
Lard	0.6	100	0	900
Margarine	0.6	81	0.4	733
Bacon, medium fat	9.1	65	1.1	626
Salad or cooking oil	0	100	0	900
Eggs				
Yolk (fresh)	16.3	31.9	0.7	355
Whole (fresh)	12.8	11.5	0.7	158
Whole (dried)	48.2	43.3	2.6	593

Table 71. Composition and Caloric Values of Selected Foods (Continued)

Food, 100 *gm.*	*Protein, gm.*	*Fat, gm.*	*Carbohydrate, gm.*	*Calories*
Vegetables				
Fresh				
Asparagus	2.2	0.2	3.9	26
Beans, lima, green	7.5	0.8	23.5	131
Beet greens	2.0	0.3	5.6	33
Cabbage	1.4	0.2	5.3	29
Carrots	1.2	0.3	9.3	45
Cauliflower	2.4	0.2	4.9	31
Corn (sweet), white or yellow	3.7	1.2	20.5	108
Lettuce	1.2	0.2	2.9	18
Onions, mature	1.4	0.2	10.3	49
Peas, green	6.7	0.4	17.7	101
Potatoes	2.0	0.1	19.1	85
Spinach	2.3	0.3	3.2	25
Sweet potatoes	1.8	0.7	27.9	125
Canned				
Asparagus	1.6	0.3	3.0	21
Carrots	0.5	0.4	6.1	30
Corn	2.0	0.5	16.1	77
Peas, green	3.4	0.4	12.9	69
Sauerkraut	1.1	0.2	3.4	20
Spinach	2.3	0.4	3.0	25
Tomato catsup	2.0	0.4	24.5	110
Tomato juice	1.8	0.2	4.3	21
Tomatoes	1.0	0.2	3.9	22
Fruit				
Fresh				
Apples	0.3	0.4	14.9	64
Apricots	1.0	0.1	12.9	56
Bananas	1.2	0.2	23.0	99
Berries	1.2	0.8	13.2	65
Blueberries	0.6	0.6	15.1	68
Strawberries	0.8	0.6	8.1	41
Grapefruit	0.5	0.2	10.1	44
Grapes	0.8	0.4	16.7	74
Lemons	0.9	0.6	8.7	44
Oranges	0.9	0.2	11.2	50
Peaches	0.5	0.1	12.0	51
Pears	0.7	0.4	15.8	70
Plums	0.7	0.2	12.9	56
Watermelons	0.5	0.2	6.9	31
Canned				
Apples; applesauce	0.2	0.1	19.7	80
Apricots	0.6	0.1	21.4	89
Orange juice	0.6	0.1	12.9	55
Peaches	0.4	0.1	18.2	75
Pears	0.2	0.1	18.4	75
Pineapples	0.4	0.1	21.1	87

Table 71. Composition and Caloric Values of Selected Foods (Continued)

Food, 100 *gm.*	*Protein, gm.*	*Fat, gm.*	*Carbohydrate, gm.*	*Calories*
Dried				
Apricots	5.2	0.4	66.9	292
Prunes	2.3	0.6	71.0	299
Raisins	2.3	0.5	71.2	298
Flour, meal				
Corn meal, whole grain	9.1	3.7	73.9	365
Cornstarch	0.5	0.2	87.0	352
Flour				
Soy, medium fat	42.5	6.5	13.6	283
Wheat, patent	10.8	0.9	75.9	355
Wheat, whole	13.0	2.0	72.4	360
Bread, white	8.5	2.0	52.3	261
Bread, whole-wheat	9.5	3.5	48.0	262
Crackers, graham	8.0	10.0	74.3	419
Corn flakes	7.9	0.7	80.3	359
Oatmeal	14.2	7.4	68.2	396
Honey	0.3	0	79.5	319
Jams, marmalades	0.5	0.3	70.8	288
Sugar, powdered or granulated	0.	0	99.5	398
Cocoa	9.0	18.8	31.0	329
Olives, green	1.5	13.5	410	144

Source: U.S. Department of Agriculture, Misc. Pub. No. 572, Washington, D.C., 1945.

water intoxication in certain conditions. An insufficient diet causes weakness and eventually death, but overeating leads to obesity and other disturbances in nutrition. There is an optimum for the total amount of food and for each one of the indispensable factors in the diet.

The dietary level of a population can be considered satisfactory when there are normal growth and development, high natality, low infantile and general mortality, a long average life span, high working efficiency, good teeth, and resistance to infection.

Caloric requirement. A technical committee of the United Nations[1] has advised the following daily caloric rations for persons living in a temperate climate:

	Kg.-cal.
Man, twenty-five years old and 65 kg. weight	3200
Woman, twenty-five years old and 55 kg. weight	2300
Woman after third month of pregnancy	Add 450
Woman during lactation	Add 1000

[1] *Ann. de la Nutrit. et de l'Aliment.*, 4, 183, 1950.

To adjust these amounts to the body weight, the following formulas are used: for men, $152 \times W^{0.73}$, and for women, $123 \times W^{0.73}$.

Sedentary men require only 2400 kg.-cal. and sedentary women 1700 kg.-cal. Hard physical labor requires an increase to 4000 kg.-cal. in man and 2800 kg.-cal. in women.

The caloric requirement of children of different ages is given in Table 72. Adolescent girls

Table 72. Caloric Need of Children

Age, years	*Kg.-cal.*	*Age, years*	*Kg.-cal.*
0.5–1	110*	10–12	2500
1–3	1200	13–15	2600†
4–6	1600	13–15	3200‡
7–9	2000		

* Per kilogram of body weight.
† Girls.
‡ Boys.

sixteen to twenty years old require 2400 kg.-cal. and boys 3800 kg.-cal., *i.e.*, 105 per cent and 120 per cent respectively of the amounts required by adult women and men.

Another estimate[1] assesses the daily caloric requirement of an adult man as 3000 kg.-cal. net, *i.e.*, the caloric value of food absorbed. Approximately 10 per cent of the caloric value is lost in preparing or cooking food or is not absorbed by the digestive tract; therefore to obtain 3000 kg.-cal. net, from 3300 to 3400 kg.-cal. gross should be allowed. An adult woman needs 2500 kg.-cal. net; *i.e.*, 2800 kg.-cal. gross. The female coefficient with respect to the caloric requirements of man is 0.83.

These figures are only averages for the total adult male population. Men leading sedentary lives need only 2400 kg.-cal. per day, while those doing heavy manual labor need 4000 to 5000 kg.-cal. Workmen in Canadian or Swedish lumber camps, who do very heavy work in a cold climate, require from 8000 to 9000 kg.-cal.

The following examples show how the caloric requirement of an individual is calculated:

An adult man 1.72 m. in height and 70 kg. in body weight has a body surface of 1.8 sq. m. His BMR is 72 kg.-cal. per hr. (using as the average standard figure 40 kg.-cal. per sq. m. per hr.). This amount covers all the caloric requirements during the 8 hr. of sleep, but during the rest of the day there is activity for which energy must be provided. Eight of the waking hours are employed in work and 8 hr. in ordinary current activities such as moving from one place to another and eating, which leads to the specific dynamic activity of food, keeping up the body temperature, etc. These activities increase the metabolic rate by approximately 30 per cent. Work requires a variable amount of energy, according to its nature; let us suppose it is covered by 1000 kg.-cal. Thus the total caloric requirement of this man is as follows:

[1] FOOD AND NUTRITION BOARD, NATIONAL RESEARCH COUNCIL, Washington, D.C., *J. A. M. A.*, **116**, 2601, 1941.

	Kg.-cal.
8 hr. of sleep at 72 kg.-cal. per hr. (BMR)	576
8 hr. awake at 94 kg.-cal. per hr. (BMR + 30 per cent)	752
8 hr. working (BMR + 1,000 kg.-cal.)	1576
Total	2904
Advisable average	3000

A woman 1.58 m. in height, weighing 56 kg., has a body surface of 1.6 sq. m. and a BMR of 37 kg.-cal. per sq. m. per hr. Energy needed for manual work can be considered as two-thirds that of a man, *i.e.*, 660 kg.-cal. Therefore the total caloric requirement is as follows:

	Kg.-cal.
8 hr. sleep at 59 kg.-cal. per hr. (BMR)	472
8 hr. awake at 77 kg.-cal. per hr. (BMR + 30 per cent)	616
8 hr. work (BMR + 660 kg.-cal.)	1132
Total	2220
Advisable average	2500

Mental activity does not increase the caloric requirement. In a cold climate the basic ration should be increased.

Children need relatively more food per kilogram of body weight than adults because (*a*) they have a higher BMR per kilogram; (*b*) part of the food is used to build up tissue; (*c*) they are in continuous activity. A pregnant woman or a nursing mother uses part of the food ingested in building up the tissues of the fetus or in producing milk for her child.

ORGANIC CONSTITUENTS OF THE DIET

The average composition of a good diet for an industrial population is given in Table 73. This diet exceeds the requirement of a man with sedentary habits and is not sufficient for one

who does hard manual labor; it is an average adequate for persons doing moderately intense manual work.

Protein. The diet of an adult should contain 1 gm. of protein per kilogram per day, which is more than the minimum protein requirement discussed in Chap. 43. Amounts varying from 20 to 130 gm. daily have been considered adequate. Undoubtedly 30 to 40 gm. of first-class protein can fulfill the minimum need of protein, but it is advisable to allow 100 or 105 gm.[1] for a diet containing 3400 kg.-cal. There should be 50 gm. of first-class protein[2] such as that in milk and milk products, eggs, meat, and glandular tissue (liver and kidney); the rest can be second-class protein, such as that of vegetables. Potato protein is one of the best of vegetable proteins. A relatively high-protein diet (1.3 to 1.5 gm. per kg.) is preferred in the United States, Western Europe, and Argentina. Countries with a low standard of living have a low-protein diet.

Table 73. Composition of Average Normal Diet

	Gm.	*Kg.-cal.*	*% of total calories*
Proteins	100	410	12
Fats	100	930	27
Carbohydrate	500	2050	61

The amount of protein in the diet should be increased during growth, pregnancy, and lactation. A pregnant woman should be given 1.5 to 2 gm. of protein per kilogram daily. The protein requirement of children varies with their age, as follows: one to three years, 3.5 gm. per kg.; three to five years, 3 gm. per kg.; five to twelve years, 2.5 gm. per kg.; fifteen to seventeen years, 2 gm. per kg.; seventeen to twenty-one years, 1.5 gm. per kg. A large proportion of the protein intake during growth, pregnancy, and lactation should be in the form of first-class protein. Muscular effort does not increase the protein requirement.

[1] A survey made in the city of Buenos Aires showed that 54 per cent of the families consumed less than 105 gm. of protein, 20 per cent consumed more; in 20 per cent of the cases the amount of protein was definitely below a safety level (ESCUDERO, P., and B. ROTHMAN, *Trab. & publ. inst. nac. nutr., Buenos Aires*, **4**, 7, 1939).

[2] These are more easily digested and absorbed and contain the essential amino acids, therefore they are more efficient in the maintenance of growth and protein equilibrium.

A greater proportion of animal protein is absorbed (90 to 95 per cent) than of vegetable protein (80 per cent). Vegetables, brown bread, cellulose, or bran in the diet decrease the absorption of animal protein.

Usually protein makes up only 12 to 15 per cent of the total caloric value of the diet, but in certain cases there is a much larger proportion, *e.g.*, the Eskimos studied by Krogh consumed 300 gm. of protein daily, their diet being almost exclusively protein and fat.

Carbohydrate. Usually more than half the calories in the diet (50 to 70 per cent) are in the form of carbohydrate. This is due to the fact that carbohydrate is cheap, agreeable, digestible, and easily absorbed. Excess carbohydrate causes flatulence owing to fermentation in the digestive tract. Carbohydrate foods are frequently refined to excess, *e.g.*, flours and sugar, and thus lose important constituents such as vitamins and protein. It is therefore advisable to replace part of the white flour and sugar in the diet by potatoes, which contain a fairly good supply of protein, thiamine, ascorbic acid, and salts.

Fat. A diet of 3000 kg.-cal. should contain from 80 to 125 gm. of fat. Usually 20 to 30 per cent of the caloric value of the diet is in the form of fat. Small children should receive approximately 4 gm. of fat per kg. The figure for adults varies considerably in different circumstances; thus Canadian lumbermen have a diet of 6000 to 8000 kg.-cal. with 300 gm. of fat (35 to 40 per cent of the total calories). It is not possible to say what amount is the indispensable minimum of fat, because rats (Mendel) and men (Pirquet, Hinhede) have been kept in good condition fed on diets with very little fat; moreover, the Japanese consume a diet which has only 20 or 30 gm. of fat daily. Nevertheless fat in adequate proportions has many advantages as a food, and a minimum is necessary to provide the indispensable linoleic, linolenic, and arachidonic acids. Also fat-soluble vitamins (A, D, E) are contained in certain fats: butter, cod- and other fish-liver oils, wheat-germ oil, etc.[1] Fats give some foods their characteristic and appetizing odor and taste. They are more satisfying than other foods and give a sensation of satiety. They are absorbed slowly, sometimes taking 5 to 6 hr. to be absorbed, so they have a prolonged nutritive effect. They contain a large

[1] Vitamin requirements are given in Table 74.

amount of calories (9.3 kg.-cal. per gm.) in a small bulk and therefore are of special value in preparing high-caloric rations.

An excess of fat, not adequately compensated by carbohydrate or protein, provokes ketosis and acidosis. Excess of fat, even when it forms part of a well-balanced diet, leads to obesity. Animal fats with high vitamin content, such as butter and fish-liver oils, are protective foods. Vegetable oils have the same caloric value as animal fats, and many have a pleasant taste, but most of them do not carry vitamins.

Table 74. Recommended Daily Dietary Allowances

	Kg.-cal.	*Protein, gm.*	*Calcium, gm.*	*Iron, mg.*	*Vitamin A, IU*	*Thiamine, mg.*	*Riboflavin, mg.*	*Niacin (nicotinic acid), mg.*	*Ascorbic acid, mg.*	*Vitamin D, IU*
Man (154 lb., 70 kg.)										
Sedentary	2,400	70	1.0	12	5000	1.2	1.8	12	75	
Physically active	3,000	70	1.0	12	5000	1.5	1.8	15	75	
With heavy work	4,500	70	1.0	12	5000	1.8	1.8	18	75	
Woman (123 lb., 56 kg.)										
Sedentary	2,000	60	1.0	12	5000	1.0	1.5	10	70	
Moderately active	2,400	60	1.0	12	5000	1.2	1.5	12	70	
Very active	3,000	60	1.0	12	5000	1.5	1.5	15	70	
Pregnancy (latter half)	2,400	85	1.5	15	6000	1.5	2.5	15	100	400
Lactation	3,000	100	2.0	15	8000	1.5	3.0	15	150	400
Children up to 12 years										
Under 1 year	110/2.2 lb. (1 kg)	3.5/2.2 lb. (1 kg)	1.0	6	1500	0.4	0.6	4	30	400
1–3 years (27 lb., 12 kg.)	1,200	40	1.0	7	2000	0.6	0.9	6	35	400
4–6 years (42 lb., 19 kg.)	1,600	50	1.0	8	2500	0.8	1.2	8	50	400
7–9 years (58 lb., 26 kg.)	2,000	60	1.0	10	3500	1.0	1.5	10	60	400
10–12 years (78 lb., 35 kg.)	2,500	70	1.2	12	4500	1.2	1.8	12	75	400
Children over 12 years										
Girls, 13–15 years (108 lb., 49 kg.)	2,600	80	1.3	15	5000	1.3	2.0	13	80	400
16–20 years (122 lb., 55 kg.)	2,400	75	1.0	15	5000	1.2	1.8	12	80	400
Boys, 13–15 years (108 lb., 49 kg.)	3,200	85	1.4	15	5000	1.5	2.0	15	90	400
16–20 years (141 lb., 64 kg.)	3,800	100	1.4	15	6000	1.7	2.5	17	100	400

Source: Food and Nutrition Board, National Research Council, Washington, D.C. (revised 1948).

MINERAL CONSTITUENTS OF THE DIET

The amount of water taken per day varies according to the diet, the nature and amount of work done, and the environmental temperature. In basal conditions it is 2,500 to 3,000 cc. per day (see Chap. 44, Water Metabolism). The importance of minerals in diet has only recently been understood. Optimum quantities of Ca and Fe are given in Table 74; the subject has been discussed in Chap. 45, Mineral Metabolism. The requirements of growing children, pregnant women, and nursing mothers are particularly important, because mineral deficiencies in the period of growth may produce effects, especially in the bones and teeth, that last throughout life.

The adult's diet should contain 0.8 gm. of calcium per day, that of women in the second half of pregnancy 1.5 gm., and that of nursing mothers 2 gm. The daily amount of phosphorus for adults should be 1.3 to 1.4 gm., and 2.5 to 3 gm. for pregnant women. A normal diet contains 3 to 6 gm. of sodium, 2 to 4 gm. of potassium, and 5 to 15 gm. of sodium chloride per day.

It is advisable that the iron content should not be less than 12 mg. per day for an adult and 15 mg. for a pregnant or adolescent woman, although 3 mg. daily suffices the needs of an adult man. The iodine content should be from 100 to 200 μg daily (0.002 to 0.004 mg. per kg. of body weight daily).

SPECIAL NUTRITIONAL NEEDS

When a diet is prescribed, age, sex, work, environmental temperature, and special conditions such as pregnancy, lactation, or disease should be taken into account.

The diet of pregnant women should be carefully controlled, because deficiencies may cause permanent injury to the fetus. Not only should extra calories be allowed, but also an adequate amount and quality of protein should be given; deficiencies in calcium, phosphorus, iron, and (in goiter districts) iodine should be carefully avoided. The following vitamins are of special importance in pregnancy: thiamine, riboflavin, niacin, ascorbic acid, and antirachitic vitamin. All these factors can be given in protective foods.

Small children especially require high-quality protein, calcium, phosphorus, and vitamins. Before 6 months, iron should be added to the diet in the form of egg yolk, strained green vegetables, or strained carrots. Potatoes are preferable to white flours.

Pregnant and nursing women should be given adequate amounts of vitamin D, especially if the exposure to the sun is restricted, either because of weather conditions or because of customs. Cod- and other fish-liver oils or vitamin D_2 concentrates are the usual forms of administering the antirachitic vitamin.

Physical exercise or manual labor does not imply an increase in protein requirement; the extra energy is obtained from carbohydrate and fat. In high-caloric diets up to 45 per cent of the caloric value may be obtained from fat.

It is preferable to use natural foods of good quality instead of covering deficiencies by adding synthetic vitamins to the diet. In certain cases, nevertheless, it is advisable to enrich foods by restoring or increasing the normal vitamin concentration of some foods. Thus up to 400 IU of vitamin D_2 can be added to 1 quart of milk with the object of controlling mild rickets. During the Second World War protective substances were added to white flours to supplement deficiencies; *e.g.*, the following quantities of vitamins and minerals were added to 1 lb. of white flour:[1] thiamine, 2.0 to 2.5 mg.; riboflavin, 1.2 to 1.5 mg.; niacin, 16 to 20 mg.; iron, 13 to 16.5 mg.; and optionally 500 to 625 mg. Ca and 250 to 1,000 IU of vitamin D. Calcium is added to counteract the effect of phytic acid in cereals, which precipitates Ca in the intestine and makes its absorption difficult. Oleomargarine should be reinforced by at least 9,000 IU of vitamin D_2 per pound.

In districts where the drinking water does not contain enough iodine to prevent endemic goiter, iodine should be added to table salt. The amount of iodine advised has varied from 5 to 200 mg. of KI per kilogram, but usually 100 mg. per kg. is considered an adequate quantity[2] (a chemical stabilizer should also be added). Each person consumes approximately 6.2 gm. of table salt daily, *i.e.*, 0.62 mg. of iodide, or 0.47 mg. of iodine; this is twice the amount usually considered as the optimum (100 to 200 mg., or 2 to 4 μg per kg., daily). In England[3] 10 mg. of KI per kilogram of salt is advised, supposing that each person consumes 10 gm. daily, *i.e.*, 100 μg of iodide or 76 μg iodine.

REMARKS ON CERTAIN FOODS

Milk. Communities where there is a high milk consumption are vigorous and have a long average life span, a high longevity rate, and a low death rate. Milk protein is of the highest biologic value, with the advantage of a low cost. It has easily digestible fat (because it is emulsified), large quantities of readily absorbable calcium and phosphorus, thiamine, riboflavin, and other vitamins. It is deficient only in vitamin D and iron. Children and pregnant women should drink 1 quart of milk daily. Where the consumption of milk is low, there are usually a low standard of living, high infant mortality, and subnormal growth and development in children. Dried, powdered milk can be used to replace ordinary fluid milk in cases where it must be transported for long distances.

Meat. Meat contains 15 to 20 per cent protein of high biologic value and 15 to 30 per cent fat, but no carbohydrate. Communities living on an almost exclusively meat diet, *e.g.*, the Eskimos and Argentine "gauchos," do not

[1] *J. A. M. A.*, **119**, 948, 1942; **122**, 437, 1943.

[2] National Committee for the Study of Endemic Goiter, *J. A. M. A.*, **121**, 423, 1943.

[3] Subcommittee in Goiter, Medical Research Council, *Lancet*, **1**, 107, 1944.

have any nutritional disorders that might be attributed to the diet such as gout, arthritis, renal disease, rickets, or scurvy.[1] A meat diet can prevent or cure scurvy or pellagra. Nevertheless from the nutritional point of view a mixed diet has many advantages over an exclusively meat diet. In a cold climate a high meat content is spontaneously chosen, possibly because its specific dynamic action contributes to maintain body heat and gives a sensation of well-being. In a hot climate, or in people who lead a sedentary life, a high meat diet is not so advantageous. The consumption of meat usually increases with the income.

Eggs. Eggs have excellent protein and a high content of iron and other minerals, vitamins A and D, thiamine, and riboflavin.

Vegetables. Green vegetables contain vitamins A, C, and B, iron, and minerals. If vegetables are boiled most of these substances are extracted and, if the water is thrown away, they are lost; for this reason steaming is preferable to boiling. Fruit contains much vitamin C, cellulose, and pectin.

Alcohol. Alcohol is oxidized in the organism, setting free 7 kg.-cal. per gm. The respiratory quotient is 0.67. As an energy-yielding food it has many disadvantages. It cannot be stored; it does not contribute to the formation of constituent substances of the body; it is toxic and habit-forming; it is expensive; it can be used as a substitute for other energy-yielding substances only for the maintenance of body temperature.

Alcohol is absorbed partly in the stomach but mainly in the small intestine. Absorption is delayed if other drink and food is ingested at the same time. The concentration of alcohol in the blood reaches a maximum approximately 1½ hr. after it has been ingested. It is oxidized in the liver (Lundsgaard; Leloir and Muñoz[2]). The concentration of alcohol in the blood falls at a steady rate related to the blood alcohol level.[3] According to Mellanby the decrease is 0.012 per cent per hour. The curve of alcohol concentration in blood is not modified by exercise.

When the concentration of alcohol in blood is 0.1 to 0.15 per 100 cc. the subject is slightly intoxicated, and definitely so when the concentration reaches 0.2 per cent except in subjects habituated to alcohol. A concentration of 0.25 per cent is usually accompanied by severe drunkenness. Little alcohol is eliminated by the kidney. The concentration of alcohol in urine is approximately 1.35 that in blood; therefore when there is 0.36 per cent alcohol in urine the subject is completely intoxicated. Alcohol is eliminated twice as rapidly by the lung as by the kidney. Thus in 8 liters of expired air there may be 3 to 4 mg. of alcohol. The condition of an alcoholized person can be estimated for purposes of medical jurisprudence by measuring the rate of alcohol elimination in the expired air.

Weak solutions of alcohol stimulate gastric secretion, but 5 to 10 per cent alcohol retards digestion, and higher concentrations have an irritating effect on the gastric mucosa, causing hypersecretion of mucus. Alcohol diminishes the absorption of vitamins in the B complex and so may lead to the production of deficiency diseases such as beriberi, polyneuritis pellagra, etc.

Alcohol increases the heart rate, causes vasodilatation in the skin, and gives a sensation of warmth. It thus favors heat loss and may have harmful results in subjects exposed to cold. It has the same effects as anoxia on the central nervous system. Persons at high altitudes (mountain climbers, aviators) should not take alcohol because the effects of alcohol are summated to those of low oxygen tension.

Kraepelin, Mott, and others have given good descriptions of the progressive effects of alcohol on the central nervous system. At first there is inhibition of the higher centers, with release of the lower centers. Critical control is diminished, pain and worry are forgotten, and there is a sensation of well-being. There are loquacity and exaggerated motility, and the subject is amenable to suggestion. During the second stage, critical control is lost, behavior is no longer normal, and emotivity is increased and uncontrolled. In a third stage, speech is slurred, vision is not clear (frequently there is diplopia), and there is incoordination of movement (alcoholic ataxia) with disturbances in posture and walking. There may be visual and auditory hallucinations. The subject is immoderately hilarious, maudlin, or quarrelsome, and in some cases immoral tendencies are no longer inhibited. The subject in this stage may become dangerous, and he is always a disgusting spectacle. In the last stages there are stupor, incontinence of sphincters, sometimes coma, and even death.

Small doses produce effects in relation to the con-

[1] Krogh, A., and M. Krogh, "A Study of the Diet and Metabolism of Eskimos," Copenhagen, 1913; Stefansson, V., *J. A. M. A.*, **87**, 25, 1926; *J. A. M. A.*, **88**, 1559, 1927.

[2] Leloir, L. F., and J. M. Muñoz, *Biochem. J.*, **32**, 299, 1938.

[3] Haggard, H. W., and L. A. Greenberg, *J. Pharmacol. & Exper. Therap.*, **52**, 150 and 167, 1934.

centration of alcohol in the blood and previous habituation. The usual effects are those of a narcotic that produces paralysis. The knee jerk decreases, sometimes after a transitory phase of exaggeration. The reaction time is lengthened, although Kraepelin found in some cases a brief period during which it was shortened. Errors increase, especially in tests requiring concentration and memory, and the subject is easily fatigued. In summary, alcohol produces an illusion of stimulation, covering disturbances in neuromotor functions and in critical control of movement and behavior; it usually diminishes efficiency in the performance of physical and intellectual tasks.

Alcohol produces an agreeable sensation of well-being and causes forgetfulness of pain and worry; hence its consumption. Taken in moderate amounts it may not be harmful, but the border line of moderation is difficult to establish and maintain. The abuse of alcohol causes many social evils apart from the harm it does to the persons indulging in it. Control of alcoholism is a major subject in social security, requiring legislation for the protection of the community and education of the people on the dangers of the abuse of alcohol.

BIBLIOGRAPHY

American Medical Association, "Handbook of Nutrition," Chicago, 1951.

Instituto Nacional de la Nutrición, Trabajos y Publicaciones, Buenos Aires.

League of Nations, publications referring to problems of food and nutrition, 1935 to 1938.

McCollum, E. V., E. Orent-Keiles, and H. G. Day, "The Newer Knowledge of Nutrition," 5th ed., Macmillan, New York, 1939.

Sherman, H. C., "Foods, Their Values and Management," Columbia University Press, New York, 1946.

United Nations, Conference on Food and Agriculture, Hot Springs, Va., 1943.

Tables of Composition of Foods

Atwater, W. O., and A. P. Bryant, "The Chemical Composition of American Food Materials," Government Printing Office, Washington, D.C., 1906.

Boyd, E. F., M. G. Eads, and H. R. Sandstead, "Food Value Tables for Calculation of Diet," U.S. Public Health Service, Washington, 1947.

Chatfield, C., and G. Adams, "Proximate Composition of American Food Materials," U.S. Department of Agriculture, Circ. No. 549, 1940.

Daniels, E. P., and H. E. Munsell, "Vitamin Content of Foods," U.S. Department of Agriculture Misc. Pub. No. 275, Washington, D.C., 1937.

Instituto Nacional de la Nutrición, "Tabla de la Composición Química de los Alimentos," Buenos Aires, 1942.

McCance, R. A., and E. M. Widdowson, "The Chemical Composition of Foods," H. M. Stationery Office, London, 1948; Chemical Publishing Company, Brooklyn, 1948.

Platt, B. S., "Tables of Representative Values of Food Commonly Used in Tropical Countries," H. M. Stationery Office, London, 1945.

Plimmer, R. H. A., "Analyses and Energy Values of Foods," H. M. Stationery Office, London, 1921.

Schall, H., "Nahrungsmittel-Tabelle," 10th ed., Verlag von Curt Kabitzsch, Leipzig, 1932.

U.S. Department of Agriculture, "Tables of Food Composition," Misc. Pub. No. 572, Washington, D.C., 1945.

Vázquez-Sánchez, J., Sobre la Composición Química de los Alimentos Españoles, *Gaceta Médica Española*, 1932 (thesis for doctorate in pharmacy).

Watt, B. K., and A. L. Merrill, "Composition of Foods, Raw, Processed, and Prepared," U. S. Department of Agriculture, Handbook No. 8, Washington, D.C., 1950.

SECTION SIX

Internal Secretions

CHAPTER 51

Introduction to Endocrinology

The glands of external secretion discharge the products of their activity on the external surface of the body, the skin, the digestive or respiratory mucosae, etc. The glands of internal secretion, also called "ductless glands" or "endocrines," pour their secretions into the blood. Certain glands are made up of two kinds of cells, some of which are exocrine and others endocrine; *e.g.*, the pancreas, the zymogenous cells of which secrete the pancreatic juice into the duodenum and the islet cells of which secrete insulin into the blood.

The glands of internal secretion produce specific chemical substances, called "hormones," which regulate important functions and are the active principles of humoral correlation between the different parts of the body. Hormonal activity was first demonstrated in 1849 by Berthold, who showed that the crest of a capon could be made to develop by grafting a testicle into the bird. The term "internal secretion" was first used by Claude Bernard (1855) when he discovered the secretion of glucose by the liver into the blood. The modern concept of internal secretion was formulated by Brown-Séquard (1869–1889) and demonstrated in 1902 by Bayliss and Starling, who showed that when acids stimulate the duodenal mucosa, a substance, secretin, is poured into the blood and is carried by the blood stream to the pancreas, where it provokes the secretion of pancreatic juice. Substances such as these, which are produced by one organ, pass into the blood, and stimulate activity in another organ, act as "chemical messengers." Bayliss and Starling called them hormones, a name suggested to them by Hardy.[1] Thus the knowledge of humoral correlation between the different parts of the organism was established. Humoral correlation appears in the embryo and in the phylogenetic scale before nervous correlation. The integration of the organism of invertebrates without a nervous system is carried out exclusively by humoral correlation; in higher animals there are mechanisms of humoral and nervous correlation, which in many cases supplement each other.

Hormones. Hormones have the following characteristics: (*a*) they are chemical entities produced specifically by a certain type of cell; (*b*) they are secreted into the *milieu intérieur* and conveyed to all parts of the body; (*c*) they provoke a specific activity in distant organs which are susceptible to their effect; (*d*) they are active in minute quantities and do not contribute energy or matter in significant amounts.

These properties distinguish hormones from other substances that are also poured into the circulating fluids, such as (*a*) nutritive substances, which provide the tissues with matter and energy, *e.g.*, glucose, amino acids, lipids, etc.; (*b*) vitamins, which are chemical regulators obtained from food; (*c*) parahormones (Gley), which are waste products of metabolism produced by all the cells of the body but which have a regulatory effect on certain functions, *e.g.*, CO_2, which regulates respiration; (*d*) chemical mediators of nerve stimulation, set free at the nerve endings on nerve cells, muscle fibers, and glandular cells, on which they have a local effect; (*e*) organizers, *i.e.*, substances in the

[1] The word "hormone" is used also for substances that do not excite activity but rather inhibit it. According to Sharpey-Schäfer, the term "autacoid" would be appropriate for all chemical messengers; "hormone" for chemical messengers that stimulate activity; and "chalone" for those which inhibit it. This terminology has not been generally adopted, chiefly because certain substances stimulate at certain doses and inhibit at others, or have an excitatory effect on some organs and an inhibitory effect on others.

embryo that exert local activity in the differentiation of tissues, even when they are transplanted to other parts of the embryo or cultivated *in vitro*.

Hormones do not initiate or suppress functions or reactions of the body; they only modify those which already exist.

According to their chemical structure hormones can be classified into three main groups: (*a*) phenol derivatives, *e.g.*, adrenaline, noradrenaline, thyroxine; (*b*) proteins, *e.g.*, the hormones of the hypophysis, pancreas, and parathyroid glands; (*c*) steroids, *e.g.*, the sexual and corticoadrenal hormones.

Active substances, known as "natural hormones," have been extracted from the ductless glands. Some of the hormones have been prepared by synthesis. Certain chemical products exert effects on the organism similar to those of hormones; they are sometimes mistakenly called "artificial hormones."

Hormones are effective in very small quantities; the minimal dose of many of them is measured in micrograms. Some of the hormones are rapidly removed from the blood, others more slowly, but in all cases hormones are being destroyed continuously, so excessive amounts are not accumulated.

Antihormones. Hormones secreted by the organism do not provoke the formation of antibodies; there is no immunity against the individual's own hormones, but extracts of ductless glands with hormonal activity, especially protein extracts of glands from other animal species, provoke the formation of antihormones which neutralize the effects of the hormone injected and may also inhibit the effect of the animal's own secretion. Hormonal substances of simple chemical structure, such as phenols (adrenaline) and steroids (ovarian, testicular, and corticoadrenal hormones), do not provoke the formation of antibodies, but protein hormones, such as the anterior hypophyseal hormones, easily provoke their formation. Insulin is a protein, but it produces antihormones in very few cases—fortunately, because insulin treatment frequently has to be kept up for many years. Prolonged treatment with parathyroid extract does not provoke the formation of an antihormone; nevertheless after a certain time the organism ceases to respond to this extract (see Functions of the pars distalis, Antihormones, in Chap. 52).

Functions of the endocrine glands. The ductless glands are of great physiologic importance, and they fulfill many functions, which may be classified as follows:

1. Metabolic, such as the increase in the rate of oxidation produced by the secretion of the thyroid, or the more specific effect of insulin on carbohydrate metabolism, or of the parathyroid hormone on calcium metabolism.
2. Morphogenetic, *i.e.*, the regulation of growth of the whole body or of certain tissues and organs, thus contributing to establish the constitution of the individual.
3. Endocrine interrelation and equilibrium, *e.g.*, the anterior hypophysis is a factor in the maintenance and development of the thyroid, adrenal cortex, and sexual glands, and reciprocally these glands regulate the activity of the anterior hypophysis.
4. Sexual and reproductive, *i.e.*, stimulation or inhibition of morphologic, functional, and psychic sexual characters and the libido.
5. Nervous and mental; hormones have considerable influence on higher nervous activity and on personality and character.
6. Resistance and adaptation, *i.e.*, they modify processes which condition the resistance, immunity, and adaptation of the body.
7. Vital; some of the ductless glands are of vital importance, so that if they are destroyed the subject dies.

At one time the ductless glands were supposed to have an antitoxic function, and endocrine insufficiencies were thought to be due to the accumulation of toxic substances that were not destroyed. These hypothetical toxic substances have never been found. On the other hand, ample proof has been given that extirpation, or destruction by disease, of an endocrine gland suppresses the secretion of a hormone that takes part in the regulation of a nutritive process; symptoms of endocrine insufficiency are due to nutritional disturbances caused by hormonal deficiency. For example, diabetes is a disturbance in carbohydrate metabolism due to insulin insufficiency, parathyroid insufficiency causes hypocalcemia, and corticoadrenal insufficiency brings about disturbances in the metabolism of salts and water. Death is caused, not by the accumulation of a toxic substance, but by these nutritive disturbances. Thus a parathyroidectomized animal can be kept alive by

the administration of calcium, and adrenalectomized animals can be treated successfully with sodium chloride. The antitoxic theory of endocrine function was definitely abandoned when the secretion of specific hormones that regulate nutritive processes was demonstrated.

Some of the hormones have a direct action on peripheral tissues; others act indirectly. Thus the pars distalis of the hypophysis secretes certain hormones which modify growth by direct action on the tissues, and others which stimulate the gonads. The sex glands thus stimulated secrete hormones which act directly on the tissues. The effect of the hypophyseal hormones on secondary sex characters is therefore exerted indirectly through the sex glands and does not take place in castrates.

The effect of one hormone can be strengthened or weakened by another hormone. Thus progesterone has a marked effect on the endometrium or on the mammary gland if these tissues have been previously developed by estrogens. In certain proportions these hormones inhibit each other.

Methods of study. The following methods are used to study the functions of the endocrine glands: (*a*) the anatomical or anatomophysiological method, which consists in the study of the macroscopic (size, weight, etc.) and microscopic (cellular changes) aspects of the gland in different physiologic and pathologic conditions; (*b*) chemical methods, which are used to extract and identify the active principles in the secretion and to study the chemical changes caused by deficiency or excess of hormone; (*c*) physiological methods, which serve to study the effects of insufficient or excessive secretion or of the injection of hormones; (*d*) the anatomoclinical method,[1] which consists in the correlation of functional disturbances observed in patients with the morphologic changes observed in the endocrine glands by means of biopsies or at autopsy.

[1] In many cases anatomoclinical observations have been the starting point of the study of problems of internal secretions, especially in the beginnings of endocrinology. Later rapid and important advances were made by the application of the experimental method, causing insufficiency by extirpation of the gland, and treating the deficiency thus created by the administration of gland extracts or (at a later stage) of purified hormones. In some cases it has not been possible to reproduce in animals the effects of the disease as observed in man.

The physiological method consists in (*a*) functional suppression of the gland and observation of the effects thus produced; (*b*) compensation of this insufficiency by glandular graft or the injection of extracts or hormones of the gland that has been removed; (*c*) the administration of excessive amounts of extracts or hormones with the object of provoking the effects of hypersecretion. The activity of a glandular extract is due in part to the specific hormone or hormones it contains, and in part to nonspecific impurities. The functions of a gland cannot therefore be deduced exclusively from the effects of its extracts. The method is nevertheless valuable when it is used simultaneously with others and gives results in agreement with them. A permanent glandular insufficiency is produced simply by removing the gland, but a permanent condition of hyperfunction is not easily maintained, because the organism tends to restore a normal functional level.

Specificity. In most cases hormonal effects are similar in different species, and a hormone obtained from one species is active in others.[1] There is, on the other hand, specificity of origin and of action, *i.e.*, a hormone is produced exclusively by one type of cell and it always has the same effect; *e.g.*, insulin is produced only by the islet cells of the pancreas and provokes hypoglycemia in all the vertebrates. Each hormone acts predominantly on one tissue or physiologic process, but it may have some effects on others. Thus testosterone exerts its principal effect on the male sexual characters, but it also has a general metabolic activity and produces some effects on the female sexual organs (endometrium, mammary gland, etc.).

Hormonal assay. Only a few hormones (*e.g.*, adrenaline) can be assayed by chemical methods. In many cases only biological methods are available. The general procedure is as follows: The different doses of the substance to be assayed are injected into a series of animals, and a definite dose of a standard preparation is injected into a second series. Certain effects, *e.g.*, the blood-sugar level, or convulsions in the case of insulin, are then observed and compared in both series. A standard preparation must be used

[1] Some of the protein hormones have a certain degree of zoological specificity. Thus sexual hypophyseal hormones (gonadotrophins) of mammals are active on the ovary of the toad only if given in large doses. Reciprocally, toad gonadotrophins have very little activity on the mammalian ovary.

because of individual variations in response, due to feeding and other causes.

The rate of secretion of a hormone can be determined in two different ways:

1. The venous blood coming from the gland is collected over a known time, and the hormone it contains is determined by chemical or biological methods. The results are usually expressed in terms of 1 gm. of glandular tissue, or 1 kg. of body weight, and 1 min. This method has given good results in estimating adrenaline secreted by the adrenal medulla, but it is not easy or even possible to apply it to other glands.
2. The dose of hormone, given in continuous or frequently repeated injections, necessary to maintain a normal functional level after removal of the gland in question, is determined. This is known as the substitution or replacement method.

An approximate idea of the rate of secretion may be obtained by measuring the hormonal concentration in the blood or the urinary elimination of the hormone, or of substances derived from the hormone. For example, the urinary elimination of 17-ketosteroids is an index of corticoadrenal and testicular activity. Similarly, pregnandiol, a derivative of progesterone, indicates the activity of the corpus luteum.

Regulation of internal secretions. Some of the ductless glands secrete continuously at a basal rate, *e.g.*, the islets of the pancreas and the adrenal medulla.[1] The rate increases or diminishes in different circumstances; *e.g.*, insulin secretion increases when there is hyperglycemia and diminishes when there is hypoglycemia.

The basal rate of secretion is maintained even after a large part of the gland has been removed; thus the blood sugar is kept at a normal level in spite of the extirpation of five-sixths to seven-eighths of the pancreatic tissue. The insufficiency of the remaining glandular tissue is made evident only in cases of emergency, *i.e.*, when there is a demand for an increase in secretion. The basal rate of secretion is not raised when the mass of glandular tissue is artificially increased. For example, three pancreases can be grafted into a dog by vascular anastomoses, thus increasing the islet tissue in the circulation to four times the normal, without provoking hypoglycemia. The total amount of insulin secreted is regulated not by the mass of pancreatic islet tissue but by the blood-sugar level. Tissue mass is important only when it is in insufficient amount.

Not only is the rate of secretion of each hormone controlled, there is also an equilibrium between the rates of secretion of the different glands on which the functional equilibrium of the whole organism is dependent.

Most of the endocrine glands continue to function normally after they have been completely denervated, because they are mainly controlled by humoral mechanisms. The adrenal medulla is an exception to this rule; after it has been completely denervated adrenaline secretion is reduced to a minimum.

Disturbances in endocrine regulation may cause (*a*) insufficiency, when the secretion is inhibited or the gland is partially destroyed and there is an insufficient mass of functioning tissue; (*b*) hypersecretion, when the nerves are stimulated (*e.g.*, sympathetic stimulation increases the rate of adrenaline secretion) or the gland is stimulated by another hormone (*e.g.*, thyroid stimulation by anterior hypophyseal thyrotrophin). Permanent hyperfunction is observed in cases of hyperplasia or adenoma of the endocrine gland. Adenomatous tissue does not respond normally to the regulatory mechanisms, but secretes an excess of hormone.

[1] Cannon has maintained there is no basal secretion of adrenaline, which is secreted only in cases of emergency.

BIBLIOGRAPHY

AMERICAN MEDICAL ASSOCIATION, "Glandular Physiology and Therapy," Chicago, 1942.

BIEDL, A., "Innere Sekretion," 4th ed., Urban & Schwarzenberg, Vienna, 1922.

EMMENS, C. W., "Hormone Assay," Academic Press, New York, 1950.

SELYE, H., "Textbook of Endocrinology," 2d ed., Acta Endocrinologica, University of Montreal, Montreal, 1949.

TURNER, C. D., "General Endocrinology," Saunders, Philadelphia, 1949.

VENNING, E. H., Assay of Hormones, *Methods in Med. Research*, **2**, 245, 1950.

Acta Endocrinologica.

Annales d'endocrinologie.

Endocrinology.

Endokrinologie.

Journal of Endocrinology.

Journal of Clinical Endocrinology.

Recent Progress in Hormone Research.

Revue française d'endocrinologie.

Vitamins & Hormones.

CHAPTER 52

The Hypophysis

THE HYPOPHYSIS IS an endocrine organ that fulfills very important functions in all vertebrates. In man it is placed in the sella turcica and measures approximately 1 cm. in its anteroposterior, 1 to 1.5 cm. in its transverse, and 0.5 cm. in its vertical diameter. It weighs from 0.5 to 0.7 gm. in man, varying with the height of the individual. It is larger in women than in men and increases in size at each pregnancy, weighing 0.8 to 0.9 gm. or even more (Comte, Erdheim, and Stumme).

The hypophysis owes its name to its situation below the brain. It was formerly called the "pituitary body," because it was mistakenly thought to secrete mucus through the cribriform plate of the ethmoid into the nasal cavities. The hypophysis is joined to the floor of the third ventricle by a thin stalk. The optic chiasma is just above the anterior part of the hypophysis, from which it is separated by the dura. The chiasma is often compressed in cases of tumor of the hypophysis.

The hypophysis is made up of two parts: (*a*) the adenohypophysis, epithelial in nature, derived from Rathke's pouch, which is a prolongation of the buccopharyngeal epithelium; (*b*) the neurohypophysis, which has its origin in an outgrowth of the floor of the third ventricle, *i.e.*, the hypothalamus. An anterior and a posterior lobe can be distinguished macroscopically, but the microscopic and embryological study of the gland has shown that it is more complex and can be divided as follows:

Adenohypophysis	Lobus glandularis	Pars distalis	Anterior lobe
		Pars tuberalis	
		Pars intermedia	Posterior lobe
Neurohypophysis	Lobus nervosus	Processus infundibularis	
	Infundibulum	Pediculus infundibularis	Hypophyseal stalk (formed by nerve fibers covered by cells of the pars tuberalis)
		Bulbus infundibularis	
		Labrum infundibularis	

The pars distalis forms the anterior lobe in mammals.[1] It is made up of nests and rows of epithelial cells separated from each other by sinusoids. There are chromophobe cells, which do not have granules, and the chromophil cells with specific granules. The granules of some of the latter stain with acid dyes; they are called acidophil, or α, cells. In others the granules stain with basic dyes; they are called basophil, or β, cells. In man 52 per cent (34 to 66 per cent) of the cells are chromophobe, 37 per cent (23 to 60 per cent) are acidophil, and 11 per cent (4 to 27 per cent) are basophil (Rasmussen). All these cells are derived from a single type of cell in the embryo, which gives rise to the different types of the adult. Chromophil cells sometimes lose their granules. There has been much discussion as to whether there is a single type of cell, which passes through different stages, or two

[1] In amphibians the pars distalis is posterior to the neurointermediate lobe.

separate types of chromophil cells. The second of these opinions is based on the following facts: (*a*) the Golgi apparatus is different in each type of cell; (*b*) hypophyseal adenomas can be made up exclusively of one type of cell, acidophil (in acromegaly) or basophil (in hypophyseal basophilism); (*c*) in different physiologic or pathologic circumstances only one type of cell may respond.

There is a cleft (the hypophyseal cleft) between the anterior and posterior lobes in nearly all mammals. Frequently a thick colloid or solid substance accumulates in this cleft. In man the cleft is found in the fetus, but not in the adult.

The pars intermedia is formed by cells of slightly basophil protoplasm, but without granules. Colloid cysts are sometimes found. The pars intermedia arises from the posterior wall of Rathke's pouch, and it does not exist in some species (fowls, whales). It is not clearly differentiated in adult man, but next to the pars nervosa there is an area in which there are colloid vesicles and basophil cells; these cells sometimes form columns which invade the pars nervosa.

The pars tuberalis surrounds the hypophyseal stalk and forms a thin layer of cells on the ventral aspect of the infundibulum. Hypophysectomy in mammals usually does not remove this part of the pars tuberalis. Atwell was the first to differentiate this part of the hypophysis in the fetus and adult. It is formed by small chromophobe cells.

The third ventricle is prolonged downward into the infundibulum; in the cat this recess extends into the pars nervosa of the hypophysis (Fig. 227). The neurohypophysis does not consist only of the processus infundibularis; it is made up of four parts: (*a*) the processus infundibularis; (*b*) the neural stalk (pediculus infundibularis); (*c*) the infundibulum (bulbus infundibularis); and (*d*) the median eminence of the tuber cinereum (labrum infundibularis). Hypophysectomy usually does not remove all the neurohypophysis; the infundibulum and the median eminence of the tuber remain.

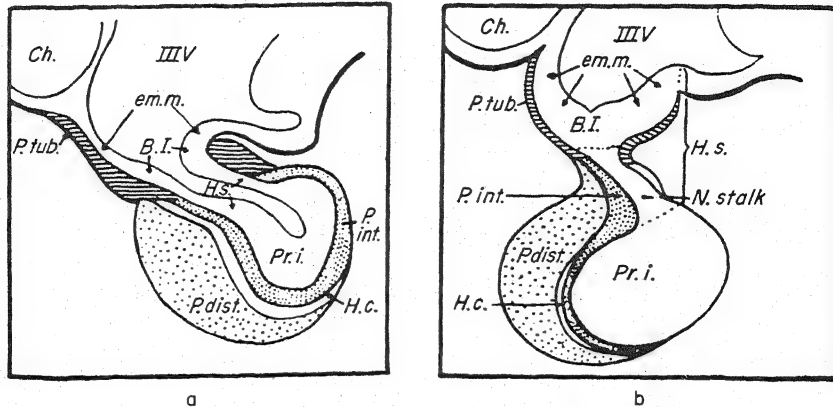



Fig. 227. Hypophysis: *a*, in the cat; *b*, in the monkey. *Ch*, optic chiasma; III*V*, third ventricle; *em.m.*, median eminence of the tuber cinereum; *P.tub.*, pars tuberalis; *B.I.*, bulbus infundibularis; *H.s.*, hypophyseal stalk; *N.stalk*, neural stalk; *P.int.*, pars intermedia; *P.dist.*, pars distalis; *Pr.i.*, processus infundibuli (pars nervosa); *H.c.*, hypophyseal cleft. (*Rioch, D., G. B. Wislocki, and J. L. O'Leary, "The Hypothalamus," Research Pub. No. XX, Association for Research in Nervous and Mental Disease, Williams & Wilkins, Baltimore, 1940.*)

The nervous lobe (processus infundibularis) is formed by neuroglia cells and by other cells with granules, known as pituicytes. There are also numerous nerve fibers.

Numerous nerve fibers from the hypothalamic nuclei, especially the supraoptic nuclei, pass through the stalk to the hypophysis (Fig. 228). Fibers from the supraoptic nuclei end in the median eminence of the tuber and in the nervous lobe of the hypophysis; a few pass into the pars intermedia and the anterior lobe. The supraoptic hypophyseal tract conducts impulses which govern the secretion of the neurohypophysis.

According to Scharrer the nerve cells of the supraoptic and paraventricular nuclei produce the posterior lobe hormones which pass along the fibers of the supraopticohypophyseal tract down to the neurohypophysis. This neurosecretion is stored there in the nerve endings from which it is released into the blood vessels. Nerve cells in these nuclei have the appearance of secretory cells (Scharrer). Gomori hematoxylin stains the neurosecretion (Bargmann). Nervous control of the anterior hypophysis through the hypophyseal stalk has been well demonstrated, but the origin and path of the impulses are not known. According to Harris this action is mediated by the blood vessels.

The hypophysis is nourished by a network of capillaries and sinusoids deriving from the carotid arteries. Popa and Fielding (1930) described a portal circulation in which blood flowed from the hypophysis to the hypothalamus. Later it was seen that blood flows from the hypothalamus to the ventral aspect of the pars distalis.[1] Damage to the vessels on the surface of the infundibulum in toads causes anemic necrotic infarcts in the pars distalis.

There is no satisfactory proof that the secretion of the hypophysis passes into the cerebrospinal fluid. The secretions of the pars distalis, the pars intermedia, and the neurohypophysis are poured into the blood stream and carried by the circulation throughout the body.

Functions of the hypophysis. The different parts of the hypophysis have the following functions:

1. *The neurohypophysis* (*a*) regulates renal excretion of water; (*b*) controls the tone of capillaries and arterioles in amphibians; (*c*) stimulates the uterus (oxytocic effect); (*d*) provokes excretion of milk.
2. *The pars intermedia* controls the dispersion of chromatosomes (pigment granules) in the pigment cells of fishes, amphibians, and reptiles.
3. *The pars distalis* regulates (*a*) growth and development; (*b*) metabolism; (*c*) endocrine functions; (*d*) sexual functions.

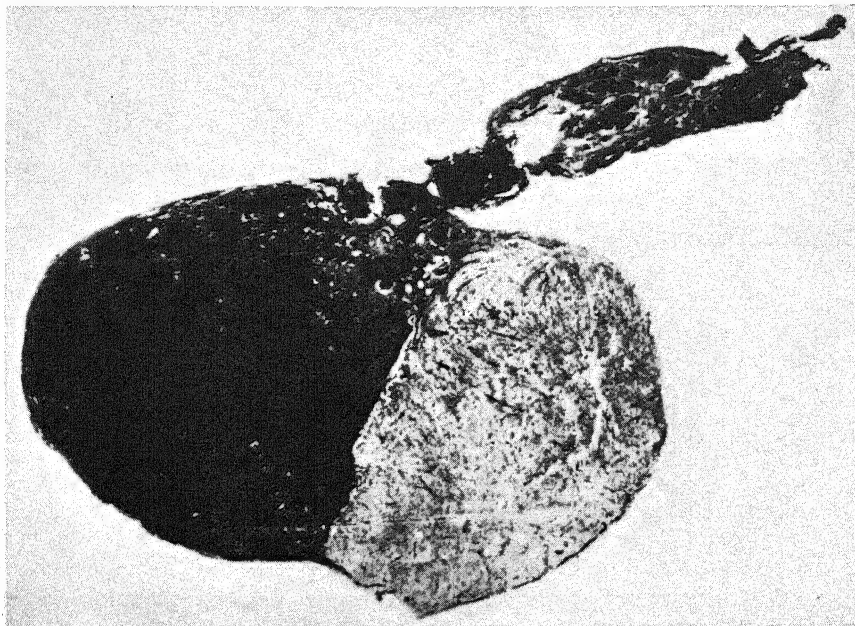



Fig. 228. Sagittal section of human hypophysis. Above, the hypophyseal stalk; to the left (dark), the pars distalis; to the right (light), the processus infundibuli.

To facilitate study, the functions of the neurohypophysis will be considered first, then those of the pars intermedia, and finally those of the pars distalis.

FUNCTIONS OF THE NEUROHYPOPHYSIS

Methods of study. Knowledge of the functions of the neurohypophysis has been obtained by (*a*) studying the effects of neurohypophyseal extracts or hormones; (*b*) partial hypophysectomy; (*c*) stimulation of the nerve centers and paths which innervate the neurohypophysis and control its function.

Hormones of the Neurohypophysis. Extracts of the neurohypophysis or the posterior lobe[1] have the following effects: (*a*) antidiuretic effect; (*b*) circulatory effect; (*c*) stimulation of contraction of the uterus (oxytocic

[1] Lascano-González, J. M., *Rev. Soc. argent. de biol.*, **11**, 309 and 318, 1935; *Compt. rend. Soc. de biol.*, **120**, 725 and 723, 1935; Green, J. D., *Anat. Rec.*, **97**, 338, 1947; **99**, 21, 1947; **100**, 273, 1948; *Am. J. Anat.*, **88**, 225, 1951.

[1] Posterior lobe extracts contain the hormones of its two parts, *i.e.*, the processus infundibularis, neural lobe, or *pars nervosa*, and the *pars intermedia* of epithelial origin.

effect) and other smooth muscles; (*d*) stimulation of excretion of milk; (*e*) other effects.

Neurohypophyseal extracts contain specific active principles, inert substances, and in some cases histamine. Kamm and his associates[1] have separated two active principles: (*a*) a pressor substance, which increases the blood pressure, has an antidiuretic effect, and causes contraction of the intestine; (*b*) an oxytocic substance, which stimulates the uterus.

Du Vigneaud and his associates have separated and purified both hormones; they have also prepared them by synthesis.[2] These hormones are polypeptides formed by nine amino acids. They differ from each other only by two of the amino acids. Two vasopressins have been found, one (in the pig) with lysine and the other (in the cow) with arginine.

The activity of posterior hypophyseal extracts is expressed in units. The international unit is the activity of 0.5 mg. of bovine posterior hypophyseal lobe extracted with acetone, dried in vacuum over P_2O_5, and kept in sealed tubes in a nitrogen atmosphere.[3] Extracts are prepared by boiling the powder for a few minutes in 0.2 per cent acetic acid solution; they usually contain 10 IU per cc.

Site of production of the active principles. The active principles are found in all parts of the neurohypophysis. At present there is no definite proof whether they are produced by the pituicytes or the nerve cells of the supraoptic nuclei. After the supraoptic hypophyseal tract has been severed in the cat, the neurohypophysis atrophies (Fig. 230); an extract of the atrophied posterior lobe does not have oxytocic, pressor, or antidiuretic effects, but still disperses the melanosomes in the pigment cells, because the pars intermedia is not atrophied.[1]

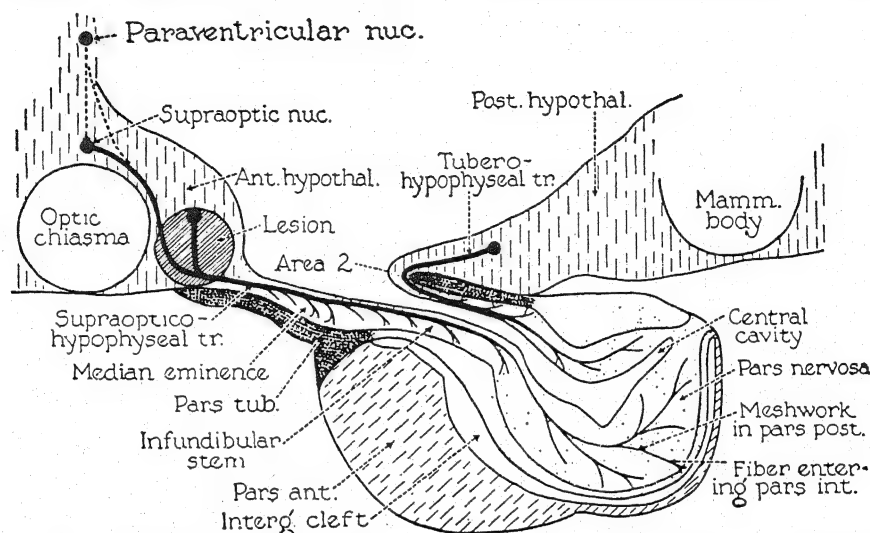



Fig. 229. Sagittal section of the hypothalamus and hypophysis in the cat, showing the hypothalamic hypophyseal and tuberohypophyseal connections. (*Fisher, C., W. R. Ingram, and S. W. Ranson, "Diabetes Insipidus and Neurohumoral Control of Water Balance," Edwards Bros., Inc., Ann Arbor, Mich., 1938.*)

Control of Renal Elimination of Water

The neurohypophysis secretes an antidiuretic hormone which stimulates reabsorption of water by the renal tubes, thus diminishing the amount of urine excreted. It is an important factor in the regulation of water metabolism.

The normal daily filtration of fluid through the glomeruli is approximately 170 liters in man. Only 1.5 liters is excreted in the urine; therefore 168.5 liters is reabsorbed by the tubes. Of the total amount reabsorbed, only one-eighth

[1] Kamm, O., *et al.*, *J. Am. Chem. Soc.*, **50**, 573, 1928; Du Vigneaud, V., *et al.*, *J. Biol. Chem.*, **186**, 77, 1950; **191**, 21, 1951.

[2] Du Vigneaud, V., *et al.*, *J. Am. Chem. Soc.*, **75**, 4879 and 4880, 1953.

[3] *Bull. Health Organ., League of Nations*, **10**, 101, 1942–1943.

[1] Fisher, C., W. R. Ingram, and S. W. Ranson, "Diabetes Insipidus and the Neurohumoral Control of Water Balance," Edwards Bros., Inc., Ann Arbor, 1938.

to one-sixth is controlled by the antidiuretic hormone (facultative reabsorption); the neurohypophysis plays no part in the reabsorption of the remainder (obligatory reabsorption).

Polyuria insipida. The neurohypophysis is innervated by fibers which arise in all the neurons in the supraoptic nuclei and the rostral part of the paraventricular nuclei (Fig. 229). The normal structure of the neurohypophysis and the secretion of the antidiuretic hormone are dependent on this innervation. Secretion of the antidiuretic hormone can be suppressed by (*a*) destruction of the supraoptic nuclei; (*b*) section of the supraopticohypophyseal tracts:[1] (*c*) extirpation of the neurohypophysis. In all these cases the operation is followed by the excretion of large amounts of dilute urine of low specific gravity. This condition is known as polyuria insipida or diabetes insipidus.[2]

There are three types of experimental diabetes insipidus:

1. When hypophyseal denervation is not complete, polyuria lasts only 3 to 5 days.
2. When denervation is complete, polyuria is established permanently.
3. In some cases there are three different stages (Fig. 231): (*a*) during an initial stage of approximately five days there is polyuria; (*b*) diuresis then returns to normal during a few days; (*c*) 14 to 15 days after the operation, polyuria is again observed and becomes permanent.

Fig. 230. Atrophy of pars nervosa after section of the supraopticohypophyseal tract. *PD*, pars distalis; *PI*, pars intermedia; *PN*, pars nervosa; *C*, central cavity of the pars nervosa; *hH*, hypophyseal cleft. (*Fisher, C., W. R. Ingram, and S. W. Ranson, "Diabetes Insipidus and Neurohumoral Control of Water Balance," Edwards Bros., Inc., Ann Arbor, Mich., 1938.*)

Hypothalamic lesions do not act directly on the kidneys through the renal nerves because they provoke polyuria in animals with denervated kidneys.[1]

During polyuria the animals show signs of intense thirst and drink continuously (polydipsia). Polyuria precedes polydipsia. Polyuria is increased by the administration of sodium chloride.

Diabetes insipidus has been observed in men and women. These patients excrete from 4 to 10 liters of dilute urine (sp. gr. 1.002 to 1.006) daily; in one case up to 43 liters was excreted. They are continuously thirsty and drink large quantities of water not only in the daytime but also during the night. If they are deprived of water, they suffer intense discomfort and their blood becomes slightly concentrated.

The pars distalis is not necessary for the production of permanent polyuria if the neurohypophysis has been completely denervated, but the excretion of urine is not so great if the pars distalis has also been removed. If part of the neurohypophysis continues to function (*e.g.*, in cases of incomplete denervation) and the pars distalis is intact, there may be polyuria, but it ceases after extirpation of the pars distalis.[2]

[1] Ranson, S. W., C. Fisher, and W. R. Ingram, The Pituitary Gland, *Research Publ., A. Nerv. & Ment. Dis.*, **17**, 410, 1938; Heinbecker, P., H. L. White, and D. Rolf, *Endocrinology*, **40**, 104, 1947; O'Connor, W. J., *Quart. J. Exper. Physiol.*, **34**, 29, 1947.

[2] The term "diabetes insipidus" was first used to distinguish this condition from the diabetes in which there is polyuria and glycosuria and the urine has a sweet taste.

[1] Houssay, B. A., and J. P. Carulla, *Compt. rend. Soc de biol.*, **83**, 1248, 1920; Baily, P., and F. Bremer, *Arch. Int. Med.*, **28**, 773, 1921.

[2] Heinbecker, P., H. L. White, and D. Rolf, *Endocrinology*, **40**, 104, 1947.

The augmentatory effect of the pars distalis on polyuria in patients with diabetes insipidus had already been reported many years ago.[1]

The pars distalis exerts its effect on diuresis through three of its hormones: the growth hormone (somatotrophin), thyrotrophin, and adrenocorticotrophin. After removal of the pars distalis, water is eliminated slowly and several renal functions are impaired.[2] Treatment with hypophyseal growth hormone compensates for these deficiencies and provokes an increase in diuresis not only in hypophysectomized animals but also in normal ones.

The effect of thyrotrophin on water excretion is mediated by the thyroid, which is stimulated by this hormone. Thyroidectomy suppresses moderate polyuria provoked by incomplete denervation of the neurohypophysis and diminishes polyuria in cases of total denervation. Thyroid treatment with doses that have no effect on normal diuresis cause an increase in polyuria insipida and provoke its appearance in latent cases.

The adrenal cortex apparently has an effect on renal excretion of water that is antagonistic to that of the antidiuretic hormone.[1] Corticoadrenal hormones control the excretion of salt and water (see Chap. 54).

There is clinical and experimental evidence that in polyuria insipida there may be a primary disturbance in the neural mechanism of thirst associated with the hormonal antidiuretic mechanism of the neurohypophysis.[2]

Neurohypophyseal extracts. Posterior hypophyseal lobe extracts and extracts of the neurohypophysis have antidiuretic activity.[3] This effect is best seen in animals which have been given water by stomach tube before injecting

FIG. 231. Polyuria and polydipsia in a typical case of experimental diabetes insipidus in the cat. *Op.*, operation; 1, permanent polyuria begins; 2, polydipsia begins. Abscissa, time in days; ordinate, urine or water in cubic centimeters. Daily urinary volume 453 cc.; daily ingestion of fluid 474 cc. (average of 6 months' observation).

[1] VON HANN, F., *Frankf. Ztschrf. Pathol.*, **21**, 337, 1918.

[2] WHITE, H. L., P. HEINBECKER, and D. ROLF, *Am. J. Physiol.*, **136**, 584, 1942.

[1] GAUNT, R., *Recent Progress in Hormone Res.*, **6**, 247, 1951.

[2] BELLOWS, R. T., and W. P. WAGENEN, *J. Nerv. & Ment. Dis.*, **88**, 417, 1938.

[3] Schaefer and Magnus (1901) and Hering (1906) first described a slight diuretic effect. Later, Farini and van der Velden (1913) discovered the antidiuretic effect.

the antidiuretic extract; the usual increase in urinary excretion is not observed, and in some cases a short period of anuria may be observed (Fig. 231). Diuresis begins again a few hours later, after the effect of the extract has passed. The antidiuretic principle acts directly on the kidney; its effects can be observed after complete renal denervation or in a perfused kidney. In patients with diabetes insipidus, subcutaneous injections of posterior lobe extracts must be repeated every few hours to maintain a normal diuresis. To avoid the discomfort of frequent injections, posterior lobe powder can be administered by aspiration through the nose, because the active principle is absorbed through the mucosa. The effect of the extract can be prolonged by combining its active principle with a substance that releases it gradually (*e.g.*, vasopressin tannate); a single injection of this preparation produces effects that last for several days.

The antidiuretic principle increases reabsorption of water from the glomerular filtrate in the renal tubes. Large doses inhibit reabsorption of NaCl. The administration of NaCl diminishes or suppresses the antidiuretic effect of posterior lobe extract. The antidiuretic principle is apparently the same substance as the pressor principle, but smaller doses are needed to obtain a decrease in the excretion of urine than to provoke a rise in blood pressure.[1]

Secretion of the antidiuretic hormone. Electrodes have been placed in the hypothalamus, and after the effects of anesthesia and trauma had passed, the nerve centers were stimulated; a decrease in urinary excretion was observed.[2] The existence of this hormone in the circulating blood is demonstrated by the following experiment: a kidney in a heart-lung preparation secretes a large amount of dilute urine; if the blood is then shunted through the head of the animal with the hypophysis intact before it passes through the kidney, the urine becomes concentrated and its volume decreases.[3]

The decrease in diuresis observed in emotional states, or provoked by reflexes or by certain drugs, has been attributed to an increase in the

[1] According to Shannon (*J. Exper. Med.*, **76**, 387, 1942) the amount secreted normally is the equivalent of 0.001 to 0.005 pressor units per hour in a 10- to 15-kg. dog (1 unit = 0.5 mg. standard posterior lobe powder).

[2] Harris, G. W., *Phil. Trans. Roy. Soc.*, **232B**, 1947.

[3] Verney, E. B., *Proc. Roy. Soc., London, s.B.*, **99**, 487, 1926; *Lancet*, **1**, 539, 1929.

secretion of the antidiuretic principle. This decrease is observed in animals with denervated kidneys, but not after suppression of the functions of the neurohypophysis.[1]

Secretion of the antidiuretic principle apparently is diminished when there is diuresis caused by ingestion of water and is increased by the injection of hypertonic salt solution. Probably in the head there are special receptors (osmoreceptors) sensitive to an increase in the concentration of the blood or tissue fluids, because injection of small amounts of hypertonic solutions of NaCl into the carotid causes a decrease in diuresis but not if the neurohypophysis has been removed or denervated.[2] Apparently the osmoreceptors are situated in the hypothalamus because local injection of hypertonic solution in anesthetized goats provokes copious drinking of water (Andersson).

An antidiuretic principle has been found in the urine of animals in anhydremia (Gilman and Goodman), which has been supposed to be the neurohypophyseal antidiuretic principle.

In amphibians, injection of the antidiuretic principle has the same effect on the kidney as in mammals. In larger doses it also increases reabsorption of water through the skin, and the animals increase in weight (Brunn).

Toxic effects. Large doses of posterior lobe extract provoke gastric ulcers and anemia (Dodds) due to hydremia and hypotonicity of the plasma. In epileptics they cause retention of water and provoke convulsions.

Action on the Uterus and in Labor

Posterior lobe extracts (or extracts of the neurohypophysis) provoke contraction of the uterus *in situ* or isolated from the body and submerged in an appropriate saline solution (Dale, 1909). The pregnant uterus is much more sensitive than the virgin uterus. The oxytocic effect is utilized in obstetrics to stimulate an inert uterus, and after birth to provoke the expulsion of the placenta or to stop uterine hemorrhage (Fig. 232). Small doses increase the tonus and rhythmic contractions of the uterus. Large doses may provoke sustained contractions. Occasionally rupture of the uterus has been observed, owing to a relatively large dose having been given to a patient with a hyperkinetic

[1] Pickford, M., *Physiol. Rev.*, **28**, 573, 1945; *Pharmacol. Rev.*, **4**, 254, 1952.

[2] Verney, E. B., *Lancet*, **2**, 739 and 781, 1946.

uterus. Purified vasopressin contains 5 to 13 per cent of the excito-uterine activity of purified oxytocin. Vasopressin is active on the human uterus.[1] Oxytocin has no pressor activity in mammals; it produces a characteristic fall in blood pressure in fowls.

Electric stimulation of the hypothalamus increases peristalsis in the rabbit (Harris). Posterior lobe extract provokes contraction of the bladder, ureter, and gall bladder, but the effect is less marked than in the uterus and intestine (Houssay, 1911). These extracts sometimes pro-

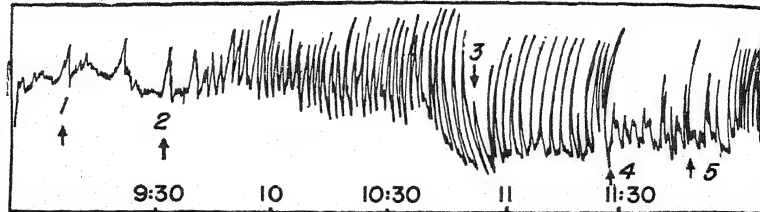



FIG. 232 Oxytocic effect of hypophyseal posterior lobe extract on the human uterus (external hysterography). 1, uterine inertia; 2, injection of posterior lobe extract; 3, expulsive contractions; 4, delivery; 5, expulsion of the placenta.

Stimulation of the hypophyseal stalk in female cats and rabbits provokes contraction of the uterus, even when the head and trunk are connected only by blood vessels.[2] Stimulation of the hypothalamus at some distance from the hypophysis provokes uterine contraction and a decrease in diuresis similar to the effects produced by injection of oxytocic posterior lobe extract.[3]

There is no definite proof that the neurohypophysis plays a part in normal labor, but the following facts are of interest in this respect: (*a*) early in pregnancy a pitocinase appears in the blood which destroys the oxytocic principle; (*b*) toward the end of pregnancy the uterus becomes sensitized toward the principle; (*c*) there is no satisfactory proof of an increase in oxytocic activity of the blood or the cerebrospinal fluid in pregnancy, or that oxytocic substances in the blood are of hypophyseal origin; (*d*) extirpation of the posterior lobe of the hypophysis (the anterior lobe remaining intact) does not prevent pregnancy and labor; (*e*) in female cats in which the neurohypophysis has been denervated (Fisher *et al.*, 1938) and in female guinea pigs with lesions in the hypothalamus behind the hypophyseal stalk (Dey *et al.*, 1941), abortion, difficult labor, or even impossibility of delivery is frequently observed.

Action on other plain muscles. Posterior-lobe extracts provoke contraction of the small and large intestine, and defecation in most subjects.

[1] SCOTT, C. R., *J. Obst. Gynaec. Brit. Emp.*, **50**, 287, 1943.

[2] HATERIUS, H. O., and J. K. W. FERGUSON, *Am. J. Physiol.*, **124**, 314, 1938; FERGUSON, J. K. W., *Surg., Gynec. & Obst.*, **73**, 259, 1941.

[3] HARRIS, G. W., *Phil. Trans. Roy. Soc.*, **232B**, 385, 1947; *Brit. M. J.*, **1**, 339, 1948.

voke bronchial constriction, but this effect is due to histamine in the extracts, not to a hypophyseal principle.

Evacuation of Milk

Extracts of the neurohypophysis or the posterior hypophyseal lobes provoke evacuation of the lactating mammary gland. The oxytocic is more potent than the vasopressor principle. Suction of the nipple produces reflex release of this hormone. Stimulation of the supraoptic nuclei or the supraopticohypophyseal tract has the same effect (see Chap. 62).

Action on the Circulation

Oliver and Schaefer discovered in 1895 that hypophyseal extracts provoke hypertension. Shortly afterward Howell (1898) showed that this effect was due to a substance in the posterior lobe.

The most characteristic effect of posterior lobe extracts, when injected intravenously, is a prolonged rise in blood pressure, which is preceded by a brief fall in blood pressure when large doses are given (Fig. 231). The rise in blood pressure lasts 10 to 20 min.; it is accompanied by bradycardia with ample beats. The change in the heart rate is considerably reduced after vagotomy and is not observed if the heart has been completely denervated before injecting the extract. The rise in blood pressure is caused by constriction of the arterioles, due to direct stimulation of the vascular muscle fibers. Constriction is also observed in isolated arterioles submerged in Ringer's fluid, when the extract is added. Coronary and

pulmonary arteries are constricted to a lesser degree and for a shorter time than arteries of other territories. Cerebral and renal arteries are not constricted; they are dilated passively by the rise in blood pressure in the intact animal injected with posterior lobe extract. A second injection does not increase the blood pressure, except after a long interval; it may even provoke a fall in blood pressure. The initial fall in blood pressure is due to a weakening of the heartbeat, which reduces the systolic volume; hypotension following a second injection of extract is caused by the same mechanism. Posterior lobe extracts are not used in therapeutics with the object of raising the blood pressure, because of their effect on the heartbeat; doses of 0.5 to 2 cc. injected subcutaneously usually provoke hypotension (a fall of 10 to 30 mm. Hg) lasting approximately 30 min. These doses cause bradycardia, preceded in some cases by transitory tachycardia, marked paleness, and a sensation of precordial constriction, sometimes extending to the limbs.

FIG. 233. Effect of hypophyseal posterior lobe extract on the blood pressure. *a*, intravenous injection of saline extract of 0.25 gm. human hypophysis into a 3.9-kg. dog anesthetized with chloralose (carotid blood pressure); *b*, intravenous injection of 0.2 gm. fresh bovine posterior lobe extract into a 17-kg. dog anesthetized with chloralose (upper tracing, arterial blood pressure; lower tracing, respiration).

The neurohypophysis is of much more importance in the maintenance of vascular tone in amphibians than in mammals. In the former, hypophysectomy causes the following disturbances: (*a*) marked capillary and arteriolar dilatation (Krogh and Rehberg, 1922; Aubrun, 1935); (*b*) an increase in capillary permeability to saline solutions, with the rapid formation of edema (Senderey); (*c*) in the toad, 40 per cent fall in blood pressure, which is not observed after removal of the pars distalis alone, or lesions in the tuber. The blood-pressure level is restored by the injection of posterior lobe extract.[1] These facts are satisfactory evidence that the neurohypophysis is an important factor in the maintenance of capillary and arteriolar tone in amphibians. Removal of the pars distalis is not followed by a fall in blood pressure until some time after the operation. Hypotension and asthenia coincide with metabolic disturbances caused by pars distalis insufficiency.

In mammals lesions of the neurohypophysis do not have such severe consequences. Section of the supraopticohypophyseal tract does not cause changes in blood pressure. Hypophysectomy produces a slight fall (20 mm. Hg) in the blood pressure of the dog[2] and rat,[3] which is apparently due to corticoadrenal insufficiency, because it is controlled by treatment with

[1] ORÍAS, O., *Rev. Soc. argent. de biol.*, **10**, 91, 1934; *Compt. rend. Soc. de biol.*, **116**, 894, 1934.

[2] BRAUN MENÉNDEZ, E., *Rev. Soc. argent. de biol.*, **8**, 463, 1932; *Compt. rend. Soc. de biol.*, **111**, 477, 1932.

[3] BRAUN MENÉNDEZ, E., and V. G. FOGLIA, *Rev. Soc. argent. de biol.*, **20**, 556, 1944.

corticoadrenal hormone or adrenocorticotrophin (Fig. 235).

Vasopressor response is subnormal in hypophysectomized animals, which are especially sensitive to hemorrhage and to drugs, such as histamine, which produce hypotension. Recovery of the blood pressure level after a relatively small hemorrhage (1.5 per cent of body weight) takes much longer than in normal animals.[1]

No satisfactory evidence of the existence in the blood of hypophyseal pressor or vasoconstrictor substances has been given. Pressor substances are found in shed blood obtained under various conditions, but they have not been proved to be produced by the hypophysis.

[1] Braun Menéndez, E., *Rev. Soc. argent. de biol.*, **10**, 204, 1934; *Compt. rend. Soc. de biol.*, **117**, 453, 1934.

Centripetal stimulation of the vagus causes a rise in blood pressure, and the oxytocic, antidiuretic, and melanosome-dispersing activities of the blood increase. These effects are not observed if the hypophysis has been removed; therefore a reflex discharge of hypophyseal secretion has been postulated.[1]

[1] Chang, H. C., *et al.*, *Chinese J. Physiol.* (several papers from 1937 to 1940).

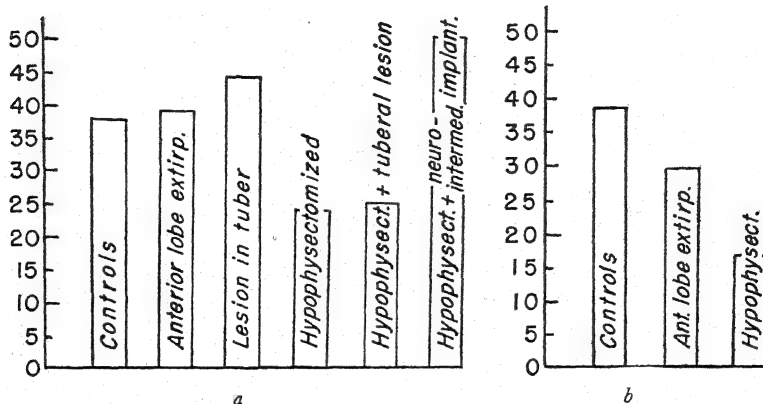



Fig. 234. Action of the hypophysis on the arterial blood pressure of *Bufo arenarum* (Hensel). Ordinate, blood pressure in mm. Hg. *a*, one week after operation; *b*, one month after operation. (*Orías, O., Rev. Soc. argent. de biol., vol. 10, p. 91, 1934.*)

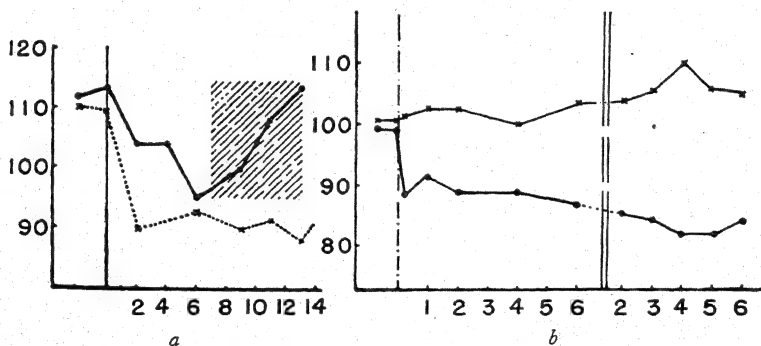



Fig. 235. Effect of adrenocorticotrophin on the arterial blood pressure of hypophysectomized rats. Ordinate, arterial blood pressure in mm. Hg. *a*, the vertical indicates hypophysectomy; abscissa to the left, time in weeks, to the right, time in days; above, blood pressure of rats which received 30 mg. daily adrenocorticotrophin (shaded area); below, controls. *b*, left vertical indicates operation; abscissa left of double vertical, time in days, right, time in weeks; above, blood pressure of craneotomized controls (average of seven rats); below, hypophysectomized rats (average of six rats). (*Braun Menéndez, E., and V. G. Foglia, Rev. Soc. argent. de biol., vol. 20, p. 556, 1944.*)

FUNCTIONS OF THE PARS INTERMEDIA

Regulation of chromatophores. The pars intermedia secretes a substance that disperses the melanosomes (pigmentary granules) in the melanocytes (pigment cells), thus conditioning to a certain extent the color of the skin in fishes, amphibians, and reptiles. This function is controlled by the nervous system. There are four

a

b c

Fig. 236. Effect of hypophysectomy on the melanocytes of *Bufo arenarum* (Hensel). *a*, left, normal toad; right, toad 10 days after hypophysectomy. *b*, section of skin of normal toad; melanosomes dispersed throughout the protoplasm and cell processes of the melanocytes. *c*, section of skin of toad 10 days after hypophysectomy; melanosomes concentrated around the nuclei of the melanocytes; the epidermal cuticle is not shed, but remains attached to the skin.

principal types of pigment cell in the skin of these animals: (*a*) melanophores, with melanin granules; (*b*) guanophores, with guanin granules; (*c*) xanthophores, with yellow granules; (*d*) erythrophores, with red granules. After total hypophysectomy, or removal of only the pars intermedia, the granules of the melanophores and erythrophores in fishes, and those of the melanophores in amphibians, are found concentrated around the nucleus; the rest of the cell, including the cellular processes, remains colorless (Fig. 236). The animals become pale, a fact which has been observed in tadpoles,[1] adult frogs,[2] and several species of fishes and reptiles.

[1] Abel, 1914; Allen, 1916; and Smith, 1916.

[2] Hogben and Winton, 1922; Giusti and Houssay, 1924.

Injection of hypophyseal extract causes intense darkening of the skin in normal and hypophysectomized amphibians.[1] This is due to dispersion of the pigment granules all over the protoplasm and cell processes, with the formation of a dark network. The cell does not expand, as was suggested by the term "melanophore-expanding substance," given to this principle. The substance should be called "melanodispersing principle," and the effects observed, "melanodispersion" and "melanoconcentration."

The pars intermedia secretes a substance that maintains a certain degree of dispersion of the melanosomes within the pigment cells. Zondek (1932) gave the name of "intermedin" to the hormone that disperses the erythrosomes in fishes, but it seems to be the same one that acts on the melanosomes of amphibians. The melanodispersing hormone exerts a continuous effect, as is shown by the rapidity with which the animals become pale after hypophysectomy. The following evidence proves that the hormone is secreted by the pars intermedia: (*a*) extracts of pars intermedia contain a higher concentration of the substance than other parts of the gland;[2] (*b*) destruction of the pars intermedia by parasites (Bayer) or by surgical removal (Atwell) in the embryo causes the development of pale animals with the pigment concentrated around the nucleus of the pigment cell, although the other parts of the hypophysis are intact; (*c*) pure cultures of cells of the pars intermedia produce the melanodispersing substance; (*d*) grafting pars intermedia into hypophysectomized tadpoles restores their normal skin color (Swingle).

The skin color of amphibians is controlled by several factors which act on the granules of the pigment cells: (*a*) the secretion of the pars intermedia; (*b*) adrenaline secreted by the adrenals; (*c*) nerves acting on the pigment cells. The hypophysis is the predominant factor in amphibians, and hypophysectomy reduces to a minimum pigmentary responses to light and darkness, and suppresses the effects of many drugs. The skin of amphibians deprived of light or blinded by removal of the eyes becomes dark[3] owing to an increase in hypophyseal secretion. When the animals are placed in a strong light they become pale because hypophyseal secretion diminishes, and the secretion of adrenaline, which has a melanoconcentrating effect, increases. Light exerts its effects on the hypophysis through the hypothalamus; section of the nerve paths from the hypothalamus to the hypophysis suppresses these effects (Stoppani).

Extracts of mammalian hypophysis also have a melanodispersing effect, but its physiologic significance is unknown. Substances that darken the skin of amphibians have been found in human urine, but it is not known whether they are produced by the hypophysis. Purified preparations of the melanodispersing hormone have been obtained by Stehle and later by Geiling.[1]

Hypophysectomy diminishes the dispersing effect of light on the melanic pigment of the retina. Injection of posterior lobe extracts restores the response to light.

FUNCTIONS OF THE PARS DISTALIS

The pars distalis, commonly called the "anterior lobe," fulfills the following functions: (*a*) stimulation of growth and development; (*b*) stimulation and control of other endocrine glands; (*c*) stimulation and control of sexual functions; (*d*) control of metabolic processes. Functions (*a*), (*b*), and (*d*) are necessary for the development and maintenance of the individual; function (*c*) is necessary for the maintenance of the species. The anterior hypophysis is the central organ of the endocrine system because it is an important factor in the development and maintenance of the structure and activity of other endocrine glands. Extirpation of the anterior hypophysis is followed by the atrophy of most of the endocrine glands. A hypophysectomized animal suffers from multiple endocrine insufficiencies and metabolic disturbances.

Methods of study. To demonstrate an anterior hypophyseal function, the following evidence must be obtained: (*a*) the removal of the anterior lobe must cause a definite and specific disturbance; (*b*) the disturbance must be controlled by restoring the anterior lobe (injection

[1] Hogben and Huxley, 1921; Houssay and Ungar, 1924–1925.

[2] In fowls, whales, and other animals without a differentiated pars intermedia, the melanodispersing substance is found in the anterior lobe, not in the pars nervosa.

[3] If the anterior half of a dark toad is joined by blood vessels to the posterior half of a light toad, the latter darkens. If the anterior half is light and the posterior half is dark, the latter becomes light. Therefore there is a hormonal factor produced in the anterior half which controls the color of the skin (Stoppani, A., *Endocrinology*, **30**, 782, 1942).

[1] Geiling, E. M. K., *Harvey Lect.*, **37**, 269, 1942; Chen, G., and E. M. K. Geiling, *J. Pharmacol. & Exper. Therap.*, **78**, 222, 1943.

of extracts or purified hormones, glandular grafts or implants); (*c*) hyperfunction obtained by implantation of the gland, or by injection of large doses of extracts or purified hormones, should produce effects opposed to those of insufficiency. All these proofs are needed for a complete demonstration; one or another alone is not sufficient evidence of the existence of the hypothetical function. In man similar evidence is obtained by (*a*) clinical, and eventually anatomical, examination of subjects suffering from diseases of the hypophysis; (*b*) treatment of the condition by injection of hypophyseal extracts or hormones, or by implantation of hormones in pellets; (*c*) surgical removal of the hypophysis or inhibition of the gland by x-rays, hormones, or other means.

Anterior hypophyseal hormones. Six hormones have been identified as secretions of the pars distalis. Usually they are called by the name of the organ or tissue or function on which they act. In some cases the suffix "-trophin" (Parkes, Corner) is added to indicate that the hormone stimulates the nutrition and functions of the cells it acts on, *e.g.*, thyrotrophin. The International Committee on Biological Assay has adopted this terminology. The suffix "-tropin," indicating that the hormone is directed to a certain tissue or organ, is also used, *e.g.*, thyrotropin. The six hormones are:

1. Growth (somatotrophic) hormone (GH), which accelerates body growth.
2. Prolactin (or luteotrophic hormone), which maintains and stimulates the secretion of the corpus luteum and stimulates the secretion of milk in mammals and the secretion of the crop glands in birds.
3. Thyrotrophin (TSH), which stimulates the thyroid gland.
4. Adrenotrophin (or corticotrophin, or adrenocorticotrophin [ACTH]), which stimulates the adrenal cortex.
5. Follicle-stimulating (or maturation) hormone (FSH), which provokes maturation of the ovarian follicle and spermatozoa.
6. Luteinizing (or interstitial-cell-stimulating) hormone (LH or ICSH), which provokes ovulation and the formation of the corpus luteum and stimulates the interstitial cells of the ovary and testicle.

These hormones are proteins; most of them have been obtained in crystallized form, and several are glucoproteins. They have been classified, according to the functions on which they exert their influence, into (*a*) metabolic hormones, and (*b*) gonadotrophins (Table 75). Some of them act directly on the tissues, *e.g.*, somatotrophin; others do so indirectly by stimulating the secretion of another gland, *e.g.*, adrenocorticotrophin stimulates the adrenal cortex, which secretes hormones acting on different tissues and functions.

Table 75. Hormones of the Pars Distalis

Function	*Hormone*	*Crystallized*	*Glucoprotein*	*Protein*
Metabolic	Somatotrophin	+	−	+
	Thyrotrophin	−	+	−
	Adrenocorticotrophin	+	−	+
Gonadotrophins	Follicle-stimulating	+	+	−
	Luteinizing	+	+	−
	Prolactin	+	−	+

Several other functions of the hypophysis and pharmacologic effects of the extracts have been observed, but the active substances responsible have not been isolated. It is not therefore justifiable to speak of hormones in these cases; it is more appropriate to refer to these as activities or factors. The main ones are (*a*) diabetogenic; (*b*) glycotropic (increase of resistance to insulin); (*c*) glycostatic (arrest of the fall in liver glycogen in hypophysectomized fasting rats); (*d*) glycogenolytic; (*e*) pancreatotrophic (hyperplasia of the islets of Langerhans); (*f*) parathyrotrophic (stimulation of the parathyroid glands); (*g*) ketogenic (increased formation of ketone bodies); (*h*) mammogenic (stimulates growth of the acini and ducts of the mammary glands); (*i*) renotrophic; (*j*) thymotrophic; etc. Purified extracts exert more than one of these effects, so probably there is not a separate principle for each one of them. Moreover, frequently more than one hormone acts on the same function.

It is not yet possible to say which of the cells of the hypophysis secrete the different hormones. There is evidence that acidophil cells secrete somatotrophin and prolactin and that basophil cells secrete gonadotrophins, thyrotrophin, and adrenocorticotrophin.

International units. International standards have been prepared and the units defined for the following hormones: (*a*) adrenocorticotrophin, 1 mg.; (*b*) prolactin, 0.1 mg.; (*c*) chorionic gonadotrophin, 0.1 mg.; (*d*) serum gonadotrophin (pregnant-mare serum), 0.25 mg.

Antihormones. After being repeatedly injected with some of the anterior hypophyseal protein hormones, animals respond less and eventually cease to respond. Globulins (antihormones) appear in the plasma of these animals which neutralize the effects of the hormone. Prolonged treatment with a hypophyseal hormone (thyrotrophin, gonadotrophin) causes atrophy of the glands involved (thyroid, gonads) of the same degree as is seen after hypophysectomy (see Antihormones, Chap. 51). The most typical and effective of the antihormones are those against thyrotrophin (antithyrotrophin) and gonadotrophins (antigonadotrophins.) The growth and diabetogenic factors also lose their potency after several injections have been given. Antihormones are specific for the hormones that they antagonize, but they are not strictly specific for the animal species that has produced them. Antihormone formation has also been observed in man; it is a serious obstacle in prolonged treatment with anterior hypophyseal hormones. In laboratory experiments it is possible to maintain the efficacy of an anterior hypophyseal preparation by periodically increasing the dose.

Animals do not produce antihormones against their own hormones. There are no antihormones normally circulating in the blood. Moreover if two animals are united in parabiosis[1] antihormones are not formed against the hormones secreted by the partners. For example, a castrate rat produces an abnormally high amount of gonadotrophin owing to hypertrophy of the anterior hypophysis; if it is united in parabiosis with a normal rat, the latter responds to the excess of gonadotrophin, and does not form antigonadotrophins.

Nervous control of the anterior hypophysis. Control of the antidiuretic secretion of the neurohypophysis by impulses arising in the supraoptic nuclei has been already discussed. Control of the secretion of the melanodispersing hormone by reflexes provoked by the action of light on the retina has also been mentioned. The pars distalis is also to a certain extent under nervous control, but the origin of the impulses and the path they follow to arrive at the hypophysis are still unknown. Moreover, few nerve fibers pass from the hypothalamus to the pars distalis. According to Harris,[1] the hypothalamus secretes substances which pass down to the hypophysis through the diencephalohypothalamic portal vessels which descend in the hypophyseal stalk and are distributed in the pars distalis. This seems to be an adrenergic effect (Markee).

Stimulation of the hypothalamus provokes secretion of gonadotrophins which cause ovulation followed by luteinization in rats and rabbits. It also provokes secretion of adrenocorticotrophin.[2]

Section of the hypophyseal stalk did not disturb the sexual cycle, ovulation, pregnancy, or lactation in some cases; in others, there were varying degrees of atrophy and hypofunction of the ovaries, testes, thyroid, and adrenals. According to Harris,[3] normal functions depended on the existence of a normal circulation between the labrum infundibularis and the hypophysis; when this circulation was not restored after section of the stalk, hypophyseal insufficiency occurred.

Certain circumscribed lesions of the hypothalamus are followed by atrophy and hypofunction of the gonads, apparently because gonadotrophin secretion is inhibited. In some cases only the secretion of the luteinizing hormone is inhibited (Dey). In others, secretion of adrenocorticotrophin caused by direct stimulation of the hypothalamus is no longer observed (Hume[4]). There is evidence that rhythmic activity of the hypothalamus controls the sexual cycle and ovulation (Everett[5]).

Intense, permanent illumination causes hypertrophy of the sexual glands and organs in certain species (ferrets, birds) during the winter season in which they are normally atrophied. Light acts on the retina and provokes reflex activity of the hypophysis; extirpation of the eyes or the hypophysis suppresses this effect of light. In species that do not ovulate spontaneously, but

[1] See page 46n.

[1] HARRIS, G. W., *Brit. M. Bull.*, **6**, 345, 1950; *Lancet*, **2**, 559 and 627, 1951; *Proc. Roy. Soc., London, s.B.*, **139**, 263, 1952.

[2] HARRIS, G. W., D. JACOBSOHN, D. M. HUME, J. DE GROOT, J. R. BROBECK, and E. EVERETT, "Ciba Foundation Symposia on Endocrinology," Vol. 4, J. & A. Churchill, London, 1951.

[3] *Ibid.*

[4] *Ibid.*

[5] *Ibid.*

only under the influence of copulation (rabbit, cat, ferret), visual stimulation (pigeons), or sexual clasping (toads), ovulation is suppressed by extirpation of the hypophysis (rabbit) or the pars distalis (toad), or by section of the hypophyseal stalk (rabbit).

Hypophyseal insufficiency. Extirpation of the hypophysis causes multiple and severe disturbances, mainly in metabolism, growth, and development. If the operation is performed in a young animal, growth is retarded or ceases, and the animal remains a dwarf with infantile characteristics. The skin remains thin and the hair is fine and soft like the lanugo of the fetus. There is hypoplasia or atrophy of the thyroid, the corticoadrenals, and the sexual glands and organs. Survival after hypophysectomy depends on the severity of metabolic disturbances, and it varies in different species. Dogs, which feed well, survive a long time if care is taken to preserve them from cold, hypoglycemia, and infection; rats, which do not eat much, lose weight and die in cachexia, after a much shorter interval; toads survive even less time.

Hypophysectomized animals must be fed frequently; otherwise the blood sugar and the liver and muscle glycogen decrease rapidly after a few hours of fasting, and they die in hypoglycemia if carbohydrate is not given in time. The BMR is low, and the body temperature is usually subnormal. The arterial blood pressure is low. Muscular tone and activity are diminished in varying degrees according to the species. If the animal is a lactating female, milk secretion ceases almost immediately after hypophysectomy and the mammary gland is rapidly atrophied. In animals that feed well (*e.g.*, the dog) there is frequently accumulation of fat which leads to moderate obesity. Hypophysectomized rats lose weight because they do not feed well. The hypophysis cannot be considered essential for life, since hypophysectomized animals survive, but they are very sensitive to cold, trauma, infections, fasting, and agents that cause hypotension or hypoglycemia. They suffer from severe metabolic disturbances which make them "frail," and great care must be taken of them or they die.

Hypophyseal insufficiency has been observed in man after surgical removal of the hypophysis, and in cases of Simmonds' disease. The latter is observed more frequently in women and is due to extensive destructive lesions of the anterior hypophysis, the most common cause of which is post-partum ischemic atrophy of the hypophysis. The principal signs are (*a*) atrophy of the gonads and sexual organs, amenorrhea, loss of libido and sexual impotence, atrophy of the mammary gland; (*b*) complete loss of pubic and axillary hair; (*c*) characteristic facies, pale and with sparse eyebrows; (*d*) dry skin; (*e*) physical and mental weakness; (*f*) low BMR, subnormal body temperature, and sensitiveness to cold; (*g*) hypersensitiveness to insulin, tendency to hypoglycemia, and low postabsorptive glycemic curve; (*h*) low urinary 17-ketosteroids. Cachexia was supposed to be a frequent outcome, but Sheehan has shown this is not so. Extreme emaciation is observed in anorexia nervosa,[1] but seldom in hypophyseal insufficiency. Autopsy shows atrophy of the endocrine glands.

Daily implantation of fresh anterior hypophysis easily corrects all the disturbances caused by hypophysectomy in the rat.[2]

Action on Growth

The pars distalis of the hypophysis is the most important endocrine factor in the stimulation and control of postnatal growth in vertebrates. Invertebrates have no hypophysis, and this gland is not necessary for the development of the vertebrate embryo and fetus; moreover growth continues in newborn mammals for several weeks after hypophysectomy. At a certain stage of development, however, the hypophysis is a necessary factor for further growth in birds and mammals. Hypophysectomy in young animals causes considerable delay, and in some cases complete cessation, of growth;[3] the animals remain dwarfs (Fig. 237). Anterior hypophyseal insufficiency in children also arrests development and they become dwarfs.

Hypophysectomized rats, forcibly fed by stomach tube to assure the ingestion of a normal

[1] In the course of anorexia nervosa some of the symptoms of hypophyseal insufficiency are observed. Prolonged undernourishment in rats causes a marked decrease in several anterior hypophyseal functions (Mulinos and Pomerantz).

[2] Treatment of hypophyseal insufficiency in man consists of injections of anterior hypophyseal extracts, which are not always efficacious. Testosterone and sometimes cortisone are also useful. Insulin injections are always dangerous because of the facility with which these patients fall into hypoglycemic shock. It is important that an adequate diet should be ingested.

[3] This was first seen by Caselli in the dog, and later was confirmed in many other vertebrate species.

amount of food, increase in weight less than normal animals; they form more fat, retain less nitrogen, and have great difficulty in forming body protein. The skeleton remains small, and bones do not grow normally, especially in length; this effect is due to a profound disturbance in the epiphyseal cartilages, where ossification is almost completely arrested. In all species the teeth grow slowly, and the eruption of the permanent teeth is retarded. In the dog the pulp cavity is large and the teeth are small (Erausquin); in the rat, dentine grows inward and obliterates the pulp cavity (Schour). All disturbances in growth caused by hypophysectomy are compensated by grafting or implanting anterior hypophyseal tissue, or injecting anterior hypophyseal extract (Smith); it is even possible by these methods to obtain growth that is greater than normal. Oral administration of anterior hypophyseal preparations has no effect in mammals.

a

b

FIG. 237. Effect of hypophysectomy on growth. *a*, four puppies from the same litter; the hypophysis had been removed from the two on the right 8 days before; *b*, three of the same animals 124 days after hypophysectomy; the dog on the left is the control.

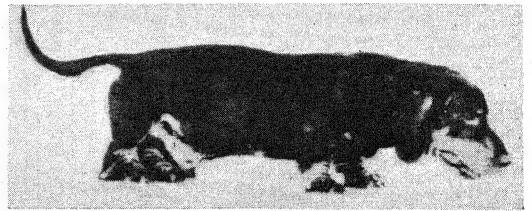
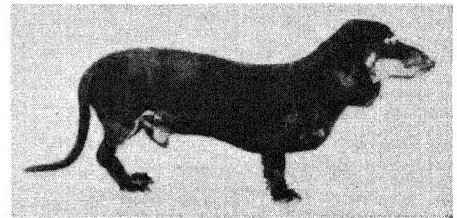
Giantism has been provoked in rats by prolonged intraperitoneal administration of anterior hypophyseal extract (Evans and Long, 1934). The weight of the injected animals is twice that of the controls. This exaggerated growth is particularly marked in the bones and viscera. Protein and water are accumulated in relatively higher proportions than fat (Lee). The epiphyseal cartilages are widened, and there is active endochondral ossification. The acceleration of growth is more marked at first in females than in males; it is especially notable in castrated females. Anterior hypophyseal extracts also accelerate growth in dogs, causing disturbances similar to those observed in human cases of acromegaly (Fig. 238).

a

b

FIG. 238. Effect of anterior hypophyseal extract on growth. *a*, dog treated with bovine anterior hypophyseal extract (10 cc. daily intraperitoneally during 6 months) shows prominent signs of giantism; *b*, litter-mate control. (*Evans, H. M., et al., "Growth and Gonad-stimulating Hormones of the Anterior Hypophysis," vol. 2, Mem., University of California, Berkeley, Calif., 1933.*)

Li and Evans[1] have obtained a purified crystalline growth hormone, which acts directly on the tissues and produces its effects not only in

[1] LI, C. H., and E. M. EVANS, *Science*, **99**, 183, 1944; *Vitamins & Hormones*, **5**, 197, 1947.

normal animals but also after hypophysectomy, thyroidectomy, or castration. Simultaneous administration of thyroid increases the effect of the growth hormone. Insulin has a favorable effect on the action of the growth hormone. Adrenocorticotrophin and several of the corticoadrenal and ovarian hormones have an unfavorable effect on the action of the growth hormone.

After an injection of growth hormone, nonprotein nitrogen decreases in blood plasma and in the liver, apparently owing to nitrogen being taken up by the tissues at a higher rate. There is a positive nitrogen balance and the amount of body protein increases as long as the animal grows. There is also retention of P, Ca, and K. P and phosphatase in plasma increase. Body fat diminishes. Liver glycogen increases, and the animals become insulin-resistant. With sufficiently high doses a diabetogenic effect is observed. If the treatment is kept up for a sufficiently long time the rate of growth diminishes in the rat; in guinea pigs growth ceases, owing to the complete ossification of the epiphyseal cartilages.

Acromegaly (Fig. 239) is a disease in which the hands and feet are abnormally developed. Its name, given by Pierre Marie, in 1866, means literally "large extremities" (Greek, ακρο-, terminal or extreme, μεγαλη, large). There is a large, prominent jaw, the zygomatic and superciliary arches and the frontal sinuses are enlarged, and the bones of the skull are irregularly thickened. The viscera are abnormally large (splanchnomegaly). The skin and the subcutaneous connective tissue are thickened. After the disease has lasted for some time there are hyperphosphatemia, a negative calcium balance, and osteoporosis of the vertebrae which leads to kyphosis. The general aspect of patients with acromegaly is striking; they have the appearance of a primitive man or gorilla, and when there is kyphosis they resemble the classical figure of Punch.

The hypophysis is hypertrophied and the sella turcica enlarged, as can be seen in radiographs of the cranium. At autopsy eosinophil adenomata or an increase in eosinophil cells in the anterior hypophysis are found. Glycosuria and diabetes have been reported in 32 per cent of the cases published. Other endocrine disturbances in acromegaly will be considered further on.

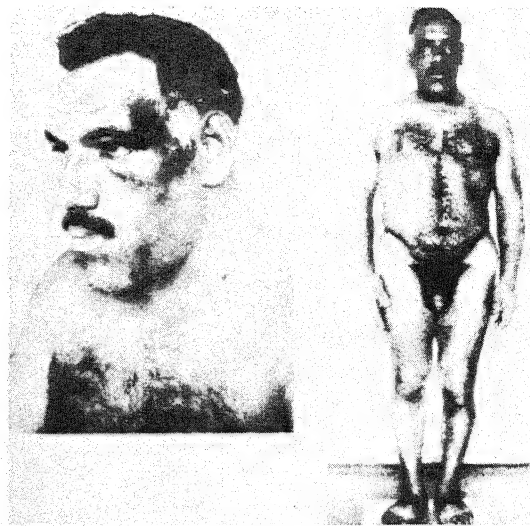
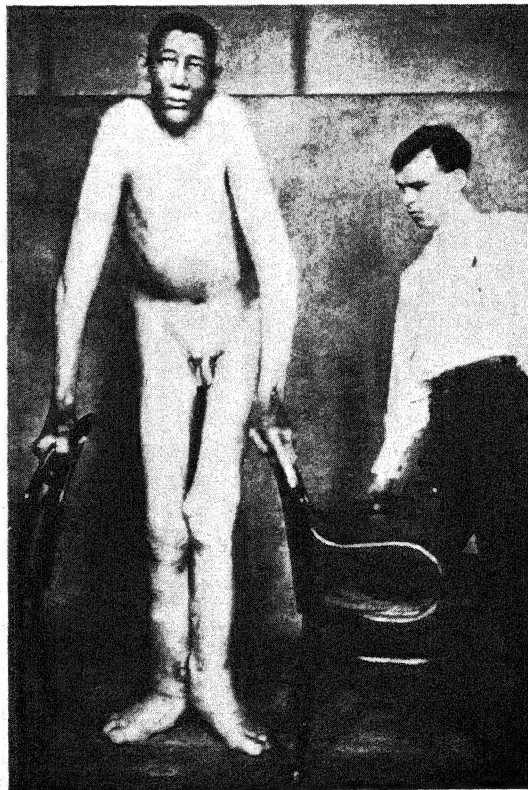
Giantism (Fig. 239) is caused by hypophyseal

a

b

FIG. 239. *a*, acromegaly. (*Courtesy of Dr. E. B. del Castillo.*) *b*, giantism; the subject's height was 2.51 m. (*Cushing, H., "The Pituitary Body and Its Disorders," Lippincott, Philadelphia, 1912.*)

hyperfunction when the cartilages are still in the process of ossification. The subject's height may be 2.30 or even 2.60 m. Signs of acromegaly are frequently found in giants.

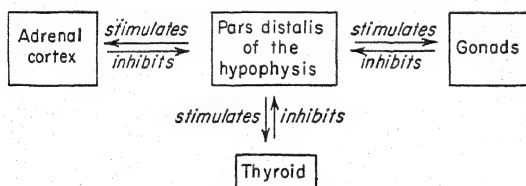
The different functions of the anterior hypophysis can be disturbed separately; thus a race of dwarf mice, which reproduces normally, has been described. There are no eosinophil cells in the hypophysis of these animals.

Hypophyseal extracts or growth hormone have not promoted growth in human dwarfs. The results cannot be compared with those obtained with gonadotrophins, testosterone, or thyroid preparations.

ACTION OF THE PARS DISTALIS ON THE DEVELOPMENT AND REGULATION OF ENDOCRINE GLANDS

The pars distalis is the central star in the endocrine constellation. It regulates other glands of internal secretion, and reciprocally, it is under the influence of certain endocrine glands. This can be summarized in the following diagram:

Adrenal cortex ⇄ (*stimulates* / *inhibits*) Pars distalis of the hypophysis ⇄ (*stimulates* / *inhibits*) Gonads

Pars distalis of the hypophysis ⇅ (*stimulates* / *inhibits*) Thyroid

The anterior hypophysis secretes hormones that stimulate the anatomical and functional development of the gonads, the thyroid, and the adrenal cortex, and which are necessary for the maintenance of the normal structure and functions of these glands. Extirpation of the anterior hypophysis in young animals causes structural and functional hypoplasia (insufficient development) of these glands; in adults it provokes atrophy and hypofunction.

Hormones secreted by the endocrine glands regulated by the anterior hypophysis have a specific effect on the structure and function of the pars distalis, as is seen in the above diagram. Castration provokes an increase in the secretion of hypophyseal gonadotrophins; adrenalectomy is followed by an increase in adrenocorticotrophin, etc. This hypophysoendocrine interrelation controls the functional state of the endocrine glands in normal and pathologic conditions.

The hormones of the anterior hypophysis act in one of three ways: (*a*) directly on the tissues, *e.g.*, the growth-stimulating factor; (*b*) by indirect stimulation, *e.g.*, gonadotrophins stimulate the ovaries, which secrete hormones stimulating the development of secondary sexual organs; (*c*) by indirect inhibition, *e.g.*, the anterior hypophyseal diabetogenic factor inhibits the pancreatic islets and suppresses insulin secretion; large amounts, injected for a sufficiently long time, damage the islet cells irreversibly and cause permanent diabetes.

Hypophysis and Thyroid

Effect of hypophysectomy on the thyroid. The pars distalis secretes thyrotrophin, a hormone that stimulates the thyroid to complete its structural and functional development. Thyroid secretion regulates the function of the pars distalis, diminishing its secretion, especially that of thyrotrophin. The hypophysis exerts a continuous effect on the thyroid; hypophysectomy therefore causes hypoplasia of the thyroid in young animals and considerable atrophy of the thyroid in adults. The same effects are observed after extirpation of the anterior lobe, but not after removal of the posterior lobe, if the pars distalis remains intact. The thyroid shows similar disturbances in human cases of hypophyseal insufficiency. The thyroid of hypophysectomized animals has a smaller volume and weighs less than that of controls. The most typical changes are flattening of the epithelium and cytological signs of hypofunction, such as a small nucleus, atrophy of the Golgi apparatus, disappearance of intracellular colloid droplets, etc. The follicles are enlarged in some species and smaller than normal in others; the colloid is not reabsorbed (Figs. 240 and 241).

The thyroid gland of hypophysectomized animals takes up less iodine than the normal, and the formation of thyroxine diminishes. Secretion of thyroxine also diminishes, and after an initial rise in blood iodine, the concentration of blood iodine diminishes and remains permanently below normal, as in cases of hypothyroidism.

There are several signs of hypofunction (not of complete functional suppression) of the thyroid in hypophysectomized animals: (*a*) BMR below normal; (*b*) low blood-iodine concentration; (*c*) increase in plasma globulins; (*d*) decrease in the phagocytic capacity of the leukocytes; etc. The thyroid is a necessary factor in the metamorphosis of amphibians, therefore if the hypophysis is extirpated in tadpoles, owing to thyroideal

hypofunction, the animals remain in the larval state. The thyroid of the axolotl always remains rudimentary, so these animals do not develop beyond the larval state, although they can reproduce. The administration of thyroid preparations, or the stimulation of the thyroid by the injection of thyrotrophin, causes metamorphosis to take place, and the animals are converted into salamanders. In adult amphibians hypophysectomy causes the normal shedding of the skin to cease, an effect also produced by thyroidectomy;[1] in birds there are disturbances in the plumage.

Effects of hypophyseal extracts on the thyroid. Anterior hypophyseal extracts have a thyrotrophic effect. A fairly pure thyrotrophin has been prepared; it has no other hypophyseal effects except a slight adrenocorticotrophic action. Thyrotrophin stimulates, anatomically and functionally, the thyroid of normal and hypophysectomized animals. Newborn chicks, guinea pigs, and toads are especially sensitive, but rats are not very sensitive to thyrotrophin.

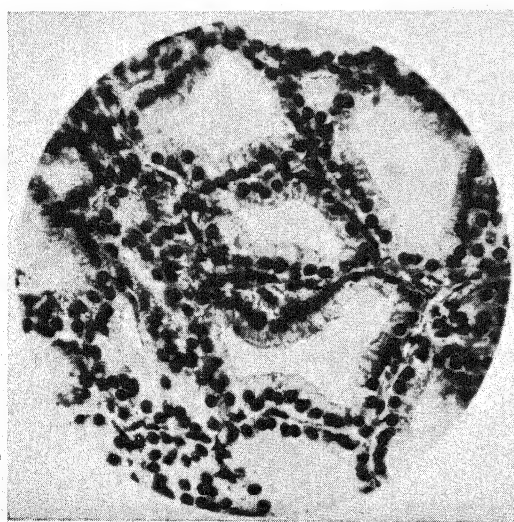
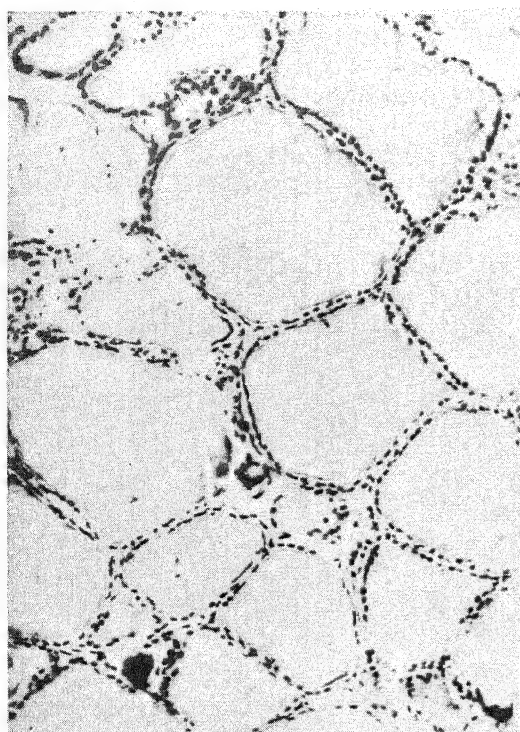
The thyroid of animals injected with thyrotrophin increases in volume, weight, and vascularization. A characteristic response is the increase in height of the epithelial cells and the appearance of intraprotoplasmic colloid (de Robertis). Many of the cells are seen undergoing mitosis[2] (Figs. 240 and 241). The colloid is reabsorbed from the follicles; its iodized substances are mobilized, and the iodine content of the thyroid diminishes, while blood iodine increases. The thyroid takes up abnormally large amounts of iodine and forms organic iodized substances at a higher rate than the normal thyroid. Thyrotrophin produces intense effects very rapidly, acting directly on the thyroid tissue. This can be demonstrated by denervating the thyroid and in thyroid grafts; or by adding thyrotrophin to slices of thyroid tissue surviving *in vitro*. Thyrotrophin is inactivated, but not destroyed, when placed in contact with thyroid, thymus, or lymphoid tissue.

Fig. 240. Effect of hypophysectomy and anterior hypophyseal extract on the thyroid of the toad *Bufo arenarum* (Hensel) Left, hypophysectomized toad; middle, normal toad; right, toad injected with extract of pars distalis.

Thyrotrophin provokes the following symptoms of hyperthyroidism in animals with an intact thyroid, which do not appear if the thyroid has been previously removed: (*a*) increase in the BMR;[1] (*b*) polyuria; (*c*) increase in blood iodine; (*d*) decrease in liver glycogen; (*e*) increase in the heart rate; (*f*) decrease in resistance to anoxia. Metamorphosis is provoked in the axolotl and accelerated in tadpoles, but not if the animals have been previously thyroidectomized. Hypophysectomized amphibians shed their skin, and there are changes in the plumage of birds.

Thyrotrophin increases the BMR in man as well as in other species. This effect is not observed in thyroidectomized subjects or in patients in whom thyroid function has been depressed by iodine treatment or other means. Thyrotrophin produces hyperthyroidism in a first stage, but later the preparation loses its efficiency, and there are signs of hypothyroidism

[1] *Bufo arenarum* (Hensel) differs from other amphibians in that hypophysectomy causes this disturbance in the skin, but thyroidectomy does not.

[2] Injection of colchicin arrests mitosis before it is completed, so the cells undergoing division can be counted. Many more cells in mitosis are found in the thyroid of animals injected with thyrotrophin than in the thyroid of the controls.

[1] Total anterior hypophyseal extract causes a slight rise in the BMR even after thyroidectomy.

such as a low BMR and flattening of the thyroid epithelium. At this stage the blood plasma has an antithyrotrophic activity.

In human cases of hyperfunction of the hypophysis (acromegaly), hypertrophy of the thyroid has been observed in 57 per cent, and a high BMR has been reported in two-thirds of the cases published (Atkinson).

Thyrotrophin is inactivated by the thyroid and eliminated in the urine in this inactive form, but it can be reactivated by heating the urine. In normal subjects active thyrotrophin is also found in the urine, but in hyperthyroid patients only the masked, inactivated form is found, because the capacity of the thyroid to inactivate thyrotrophin is increased in these cases. On the contrary, in patients with thyroid insufficiency only free thyrotrophin is found in urine.[1]

Thyrotrophin causes exophthalmos (protruding eyeballs), which is more marked in thyroidectomized animals than in those with an intact thyroid.[2] According to Smelser, exophthalmos is produced in this case by edema of the tissues of the orbit; Dobyns considers that it is due mainly to an increase in the fatty tissues of the orbit.

The hypophysis of thyroidectomized animals. The thyroid has a definite influence on the pars distalis of the hypophysis. Thyroidectomy causes an increase in the volume and weight of the anterior lobe of the hypophysis, which on histologic examination is seen to contain considerably fewer eosinophil cells and more basophil cells than the normal hypophysis. Numerous fine colloid vacuoles, coalescing to form larger ones, are observed in the basophil cells, and in some instances a single very large vacuole.[3]

Hypophyseal colloid increases after thyroidectomy. This colloid substance differs from that in the thyroid by its lower iodine content and several other features.

The pars distalis of thyroidectomized animals

FIG. 241. Effect of hypophysectomy and anterior hypophyseal extract on the thyroid of the dog. Above, hypophysectomized dog; the epithelium is flat and atrophic. Below, animal injected with anterior hypophyseal extract; the epithelium is high and forms folds, the colloid is fluid and vacuolated.

[1] MEANS, J. H., *Am. J. M. Sc.*, **1**, 207, 1944.

[2] Schockaert in ducks; Marine and Rosen in guinea pigs.

[3] Castration also causes the appearance of vacuolated cells in the anterior hypophysis, but they can be differentiated by cytological methods from those seen after thyroidectomy. Thyroid treatment is more efficacious than the injection of sexual hormones for the normalization of hypophyseal cells of thyroidectomized animals, and sexual hormones are more efficacious than thyroid preparations for the normalization of hypophyseal "castration" cells.

has less thyrotrophin[1] and gonadotrophin[2] than is found in the hypophysis of normal rats.

Changes in the hypophysis produced by thyroidectomy are not compensatory responses to thyroid insufficiency, because they do not diminish the symptoms of hypothyroidism. Thus hypophysectomy following thyroidectomy does not cause a further fall in the already low BMR.

The effect of thyroid treatment on the hypophysis. Thyroid treatment restores the normal aspect of the anterior hypophysis in thyroidectomized animals. An excess of thyroid causes an increase in the number of eosinophil cells, a small increase in basophil cells, and reabsorption of the hypophyseal colloid. Thyrotrophin activity of the anterior hypophysis is considerably diminished and disappears completely if the treatment is sufficiently intense and prolonged. The sexual activity of the hypophysis increases in animals on thyroid treatment, if the treatment is not too intense (Reforzo-Membrives), but if thyroid is given in large doses over a long period, the gonadotrophin content of the anterior hypophysis diminishes (Arrighi).

Adrenotrophic Action of the Hypophysis

The pars distalis of the hypophysis secretes a substance which continuously stimulates the adrenal cortex and maintains its normal structure and function.[3] This substance is known as adrenocorticotrophin or adrenocorticotrophic hormone (ACTH).

Effect of hypophysectomy on the adrenals. Extirpation of the hypophysis, or of its anterior lobe, causes atrophy of the adrenal cortex (Fig. 242). No structural changes are seen in the adrenal medulla and its adrenaline content is not altered. Corticoadrenal atrophy has also been observed in human cases of hypophyseal insufficiency. Retrogressive changes are especially marked in the central part of the adrenal cortex, *i.e.*, the reticular zone and the deeper part of the fascicular zone. Compensatory hypertrophy of the adrenal cortex is not observed in hypophysectomized animals. Thus if subtotal adrenalectomy is performed in normal animals the remaining cortical tissue hypertrophies, but this does not occur in hypophysectomized animals; therefore adreno-

Fig. 242. Effect of hypophysectomy on the adrenal. Above, transverse section of normal rat adrenal. Below, transverse section of the adrenal from a rat 1 month after hypophysectomy. The medulla is normal; the cortex is atrophic, especially the inner layers (light); the outer layers appear dark because the cells have little protoplasm and the nuclei of the different cells are close to each other.

[1] If a thyroidectomized rat is joined in parabiosis with a normal one, there is no increase in the weight or activity of the normal animal's thyroid.

[2] Reforzo-Membrives, J., *Endocrinology*, **32**, 263, 1943.

[3] Houssay, B. A., *Prensa méd. argent.*, **20**, 1563, 1933; Swann, H. G., *Physiol. Rev.*, **20**, 493, 1940; Tepperman, J., F. L. Engel, and C. N. H. Long, *Endocrinology*, **32**, 373, 1943.

corticotrophin is a necessary factor in this process. The cytological process of atrophy varies in different species. In hypophysectomized rats the corticoadrenal cells are small and have no fatty droplets; there is a light zone near the medulla, where many of the cells have disappeared, surrounded by a zone that appears dark because it is made up mostly of nuclei, owing to the cells having lost most of their protoplasm. In dogs and toads the cells are at first overloaded with large droplets of fat; later, in the dog, there is vacuolar or areolar degeneration of the protoplasm, and the cells die gradually.

The adrenal cortex of hypophysectomized animals has the structural appearance of hypofunction, but it continues to secrete enough hormones to maintain the animals alive. Adrenalectomy is followed by a typical and fatal syndrome of adrenal insufficiency.

Adrenocorticotrophin. Injection of anterior hypophyseal extract or of ACTH stimulates the adrenal cortex. In hypophysectomized animals it prevents adrenal atrophy, or restores the normal condition of the adrenals if they are atrophied. In normal animals it causes adrenal hypertrophy.

This hormone was separated as an apparently pure thermolabile protein;[1] later, by electrophoresis or by adsorption on oxycellulose, a preparation 100 times more potent was obtained.[2] Active polypeptides have been obtained by hydrolysis. Probably ACTH is a complex substance, because parts with different properties have been split off the parent substance.

Adrenal hyperfunction produced by injection of ACTH or by an increase in the secretion of the hormone by the pars distalis is revealed by the following signs, which are not seen in adrenalectomized animals:

1. Atrophy of the thymus and lymphoid tissue.
2. Decrease in lymphocytes and eosinophil cells in the blood. Since eosinopenia does not occur if the adrenal is not functioning, it is used as a clinical test of adrenal function (Thorn).
3. Increase in blood sugar, liver glycogen, and resistance to fatigue in hypophysectomized animals.
4. Increase in protein catabolism, especially if large doses are given.
5. Increased urinary elimination of uric acid in man.
6. Increased urinary elimination of 17-ketosteroids, oxysteroids, and glucocorticoids.
7. Retention of water, Cl, and Na and increased elimination of K.
8. Increase in the severity of diabetes that has been attenuated by hypophysectomy (Long). In man occasionally diabetic signs have been observed (Conn *et al.*).
9. Increased resistance to insulin.
10. Increase in liver arginase.

In some species it retards ossification and general growth, behaving in this respect as an antagonist of the growth hormone (see page 622).

ACTH causes marked and rapid decrease in the ascorbic-acid and cholesterol content of the adrenal cortex. ACTH can be rapidly and accurately assayed by injecting it into a hypophysectomized rat and measuring the decrease in the ascorbic acid of the adrenal cortex provoked in 1 hr.[1]

There is a rapid increase in the secretion of ACTH, which stimulates the adrenal cortex in many circumstances in which the organism is under stress, such as cold, trauma, violent exercise, or the administration of certain drugs (histamine, formaldehyde, etc.).

The adrenal gland normally inhibits the hypophyseal secretion of ACTH. Adrenalectomy is followed by an increase in ACTH secreted into the blood.

In cases of hyperpituitarism in man, an increase in the size of the adrenals is frequently observed, and in some cases of acromegaly adenomas made up of corticoadrenal cells have been found (see Cushing's disease, Chap. 54).

The Hypophysis and the Sexual Glands

The anterior hypophysis secretes gonadotrophins, which contribute to the development and maintenance of the structure and function of the sexual glands and organs and play a part in

[1] Li, C. H., H. M. Evans, and M. E. Simpson, *J. Biol. Chem.*, **149**, 413, 1943; Sayers, G., A. White, and C. N. H. Long, *J. Biol. Chem.* **149**, 425, 1943.

[2] Astwood, E. B., M. S. Raben, R. W. Payne, and A. B. Grady, *J. Am. Chem. Soc.*, **73**, 6, 1951; Dixon, H. B. F., *et al.*, *Nature, London*, **168**, 1044 and 1084, 1951.

[1] Sayers, M. A., G. Sayers, and L. A. Woodbury, *Endocrinology*, **42**, 379, 1948.

the regulation of the sexual cycle, estrus and menstruation, ovulation, libido, pregnancy, and lactation. Reciprocally the hormones of the ovary and testes regulate hypophyseal structure and function. These facts will be considered when dealing with reproduction in Sec. VII.

Interrelations between the Hypophysis and other Endocrine Glands

Parathyroids. There is no definite proof that the anterior lobe of the hypophysis secretes a hormone that contributes to the development and maintenance of the parathyroid glands, but hypophysectomy is followed by certain changes in these glands. Thus they are smaller in hypophysectomized than in normal tadpoles (Smith) and are made up of abnormally small cells in the rat and monkey. Retrogressive lesions have been observed in the parathyroids of 60 per cent of hypophysectomized dogs and in only 10 per cent of the controls (Houssay and Sammartino). These lesions were seen only in limited areas. They consisted in accumulations of nuclei surrounded by very little protoplasm (syncytioid structures) and in hyaline areas deprived of cells. In hypophysectomized pancreatectomized dogs these lesions were more marked and frequent. They may be due to general nutritive disturbances and not to the absence of a specific parathyrotrophic hormone. Moreover hypophysectomized dogs have a normal blood-calcium level (Marenzi and Gerschman), therefore there cannot be a marked condition of parathyroid insufficiency. Anterior hypophyseal extracts have been reported to produce histologic signs of hyperactivity in the parathyroid with an increase in calcemia (Anselmino and Hoffman), which does not occur in parathyroidectomized animals. In only a few cases of acromegaly or hypophyseal basophilism has the condition of the parathyroids been reported; in some they were found to be normal; in others hyperplasia and in a few cases adenomas were found.

Pancreas. The anterior hypophysis has no pancreatotrophic effect. Hypophysectomy is not followed by atrophy of the pancreatic islets. In hypophysectomized animals the zymogenous and endocrine tissue of the pancreas is less developed than in normal animals, but since the deficiency in zymogenous tissue is greater, there is more islet tissue per unit volume or weight of pancreas. The islet structure, insulin content, and insulin secretion are normal.[1] Anterior hypophyseal extracts produce two types of response according to the dose, the animal, and the condition of the pancreatic islets: (*a*) an increase in the islets and insulin content, which Anselmino and Hoffman called the "pancreatotropic effect," attributing it to a specific hormone; (*b*) lesions in the islet cells, which are transitory or permanent, causing hypophyseal or metahypophyseal diabetes (see Chap. 41).

METABOLIC FUNCTIONS OF THE PARS DISTALIS

The pars distalis of the hypophysis has a regulatory effect on many metabolic processes. Some of these effects are exerted directly by the hypophysis, others indirectly through stimulation of another gland. These secondary effects may be of a general character, *e.g.*, the action of the thyroid on oxygen consumption by the tissues. Others are localized to a certain tissue, *e.g.*, the effects of ovarian hormones on the uterus.

Basal Metabolism

The pars distalis of the hypophysis has only a slight direct effect on the BMR; most of its action is exerted through the thyroid gland. Hypophysectomy causes atrophy of the thyroid and a decrease in the BMR, which is on an average −16 per cent in the dog, −25 to −40 per cent in the rat, and considerably below normal in patients suffering from hypophyseal cachexia. Thyroid function is depressed but not completely abolished, because thyroidectomy causes a further decrease in the BMR to the level usually found in thyroidectomized animals. The administration of anterior hypophyseal extract, or of thyrotrophin, causes hyperplasia of the thyroid and an increase in the BMR; but in thyroidectomized animals the same treatment produces only a small and transitory rise in BMR.

Specific dynamic action (SDA). Hypophysectomy does not modify the specific dynamic action of protein in dogs (Mazzocco, Artundo). In man it usually causes a decrease, but there is

[1] Houssay, B. A., *Medicina, Buenos Aires*, **2**, 205, 1941; Houssay, B. A., V. G. Foglia, F. S. Smyth, C. T. Rietti, and A. B. Houssay, *Rev. Soc. argent. de biol.*, **17**, 301, 1941; *J. Exper. Med.*, **75**, 547, 1942.

no strict correlation between the variations in the SDA and the symptoms of insufficiency or excessive function of the hypophysis. The diagnostic significance of a decrease in the specific dynamic action, as a sign of hypophyseal insufficiency, has been overrated, especially as many extrahypophyseal factors cause a decrease in SDA.

Protein Metabolism

The pars distalis is one of the main factors in the regulation of protein synthesis in the intact organism of mammals. The effect of the hypophysis on postnatal growth is due principally to its action on protein synthesis.

Hypophyseal insufficiency. In hypophyseal insufficiency growth is retarded, less nitrogen is retained, and less body protein is formed than in normal conditions. Hypophysectomized animals rapidly catabolize ingested protein and the urinary elimination of nitrogen is normal, but in emergencies the ability to mobilize and consume body protein is below normal. Hypophysectomized dogs excrete less nitrogen than the normal control during total or protein fasting (Aschner, Braier). Hypophysectomized rats, which eat little, lose weight rapidly and fall into cachexia; they excrete more nitrogen than the controls. In pancreatic or phlorhizin diabetes, hypophysectomized animals excrete less nitrogen, lose less weight, and convert less protein into glucose than the controls (Houssay and Biasotti).

The effect of anterior hypophyseal extracts. Injection of anterior hypophyseal extract, or better still of growth hormone, causes a decrease in the concentration of blood nonprotein nitrogen and in the urinary excretion of nitrogen in normal fed or fasting animals and in phlorhizinized animals. This seems to indicate that the tissues take up nitrogen at an increased rate. Repeated and prolonged administration of the extract provokes growth with the formation of body protein and a positive nitrogen balance.

Fat Metabolism

Hypophyseal insufficiency. Moderate obesity is sometimes observed in hypophysectomized dogs. Hypophysectomized rats and patients with hypophyseal insufficiency, on the contrary, have a poor appetite, lose weight rapidly, and fall into cachexia. Forcible feeding by stomach tube causes hypophysectomized rats to increase in weight; they form less body protein and grow less than the controls, but they accumulate much more fat.

Lesions in the hypothalamus of normal or hypophysectomized animals provoke obesity of enormous proportions. There is always a marked increase in appetite, and overeating seems to be the main cause of this type of obesity, but metabolic disturbances are also present.

Hypophysectomized dogs catabolize less fat and have a lower basal urinary excretion of ketone bodies. In pancreatic and phlorhizin diabetes and in fasting, ketonuria does not increase much in hypophysectomized animals (Rietti). These facts show that the capacity of these animals to mobilize and catabolize fat in emergencies is diminished.

Effect of anterior hypophyseal extract. Extracts of the anterior lobe of the hypophysis provoke a marked increase in blood-fat, and in the fat content of the liver and pancreas, with a decrease in muscle- and bone-fat. Ketonemia and ketonuria increase, indicating an increased rate of fat catabolism; but this is not observed in hepatectomized animals. Somatotrophin and adrenocorticotrophin increase liver-fat, ketonemia, and ketonuria. The hypophyseal growth hormone increases body protein and water, but diminishes the fat content of the body. ACTH sometimes increases body-fat. The anterior hypophysis, therefore, controls fat mobilization and consumption.

Carbohydrate Metabolism

The important part played by the hypophysis in carbohydrate metabolism has already been discussed (see Chap. 41).

Mineral Metabolism

Hypophysectomized animals retain less mineral elements because they form less body substance owing to the arrested growth. On the contrary hypophyseal extracts increase the retention of phosphorus, calcium, potassium, and other elements that are constituents of the tissues. ACTH provokes increased excretion of nitrogen, phosphorus, and potassium, especially when it has marked lympholytic effect. Anterior lobe extracts increase diuresis because they provoke hyperthyroidism; they do not have this effect in thyroidectomized animals.

BIBLIOGRAPHY

Annual Review of Biochemistry.

Annual Review of Physiology.

Association for Research in Nervous and Mental Diseases, "The Pituitary Body" (Research Pub. No. XVII; several papers by 46 authors), Williams & Wilkins, Baltimore, 1938.

Association for Research in Nervous and Mental Diseases, "The Hypothalamus" (Research Pub. No. XX; several papers by 34 authors), Williams & Wilkins, Baltimore, 1940.

Houssay, B. A., "Advancement of Knowledge of the Rôle of the Hypophysis in Carbohydrate Metabolism during the Last Twenty-five Years," *Endocrinol.*, **30**, 884, 1942.

Houssay, B. A., "La Acción Fisiológica de los Extractos Hipofisarios," 2d ed., Flaiban, Buenos Aires, 1922.

Houssay, H. E. J., "Hipófisis y Crecimiento," El Ateneo, Buenos Aires, 1947.

Li, C. H., Growth and Adrenocorticotrophic Hormones of the Anterior Pituitary, *Harvey Lect.*, New York, 1951.

Long, C. N. H., The Influence of the Pituitary and Adrenal Glands upon Pancreatic Diabetes, *Medicine*, **16**, 215, 1937.

Long, C. N. H., A Discussion on the Mechanism of Action of Adrenal Cortical Hormones on Carbohydrate and Protein Metabolism, *Endocrinology*, **30**, 870, 1942.

Long, C. N. H., *Ann. New York Acad. Sc.*, **43**, 383, 1943.

Pickford, M., Control of the Secretion of the Antidiuretic Hormone from the *Pars Nervosa* of the Pituitary, *Physiol. Rev.*, **25**, 573, 1945.

Pickford, M., Antidiuretic Hormone Secretion, *Pharmacol. Rev.*, **4**, 254, 1952.

Russell, J. A., The Relation of the Anterior Pituitary to Carbohydrate Metabolism, *Physiol. Rev.*, **18**, 1, 1938.

Severinghaus, A. E., Cellular Changes in the Anterior Hypophysis with Special Reference to its Secretory Activities, *Physiol. Rev.*, **17**, 556, 1937.

Sheehan, H. L., *et al.*, *Quart. J. Med.*, **32**, 277, 1939; **42**, 319, 1949 (human hypopituitarism).

Stehle, R. L., *Vitamins & Hormones*, **7**, 389, 1949; **8**, 215, 1950 (posterior lobe hormones).

Van Dyke, H. B., "The Physiology and Pharmacology of the Pituitary Body," two vols., University of Chicago Press, Chicago, 1936 and 1939.

White, A., Luteotrophin (prolactin), *Vitamins & Hormones*, **7**, 253, 1949.

CHAPTER 53

The Thyroid Gland

Comparative anatomy and embryology. The thyroid first appears in vertebrates as a gland that pours its secretion into the digestive tract. Later the glandular duct is closed and the thyroid becomes an endocrine gland, secreting an iodine-containing substance, which is an important factor in the regulation of metabolism, growth, and morphogenesis. In the human embryo the gland is formed by an outgrowth of the ventral aspect of the pharynx at the level of the first branchial cleft, and two lateral growths originating in the fourth branchial cleft. A duct (the thyroglossal duct) is formed which is later closed up and reabsorbed. The foramen caecum at the root of the tongue is a vestigial remainder of this duct. Small thyroid growths are occasionally found in the path of the fetal thyroglossal duct, which sometimes hypertrophy and become goitrous.

Physiologic anatomy. The thyroid of a normal adult weighs from 20 to 40 gm. (approximately 0.4 gm. per kg. of body weight), although in certain geographical areas it is found to be larger. It varies considerably in size, and responds with structural and functional changes to different internal or external factors. The thyroid shows signs of activity in the fetus, but it does not reach its maximum functional development until puberty. In old age it retrogresses. In the female it is larger than in the male and is subiect to changes coinciding with the sexual cycle, pregnancy, lactation, and menopause. Diseases of the thyroid are more frequently observed in women than in men. In certain species the thyroid develops considerable activity (*e.g.*, the rat); in others its functions are never fully developed (*e.g.*, the axolotl).

There is a definite seasonal variation in the thyroid. At the end of winter and the beginning of spring it is larger and its iodine content is lower than in summer. The highest iodine content and lowest weight are found toward the end of summer and the beginning of the fall. In goitrous areas the incidence of goiter (enlargement of the thyroid) increases when the iodine content is lowest, *i.e.*, at the end of winter.

The thyroid, like other endocrine glands, has a very active circulation. The minute volume (3.5 to 6.9 cc. per gm.) is greater than that of the brain or the kidney. It has also a rich lymphatic network. It is innervated by fibers of the sympathetic arising in the first six or seven thoracic segments, and of the parasympathetic (laryngeal nerves). There are fibers with vasomotor activity. The existence of secretory fibers has been postulated (Asher, Cannon) but has not been satisfactorily demonstrated, and it is a well-proved fact that the thyroid functions and responds to different stimuli after complete denervation or when grafted (*i.e.*, when it has no nervous connections). Several clinical and experimental observations show that the central nervous system exerts an influence on the thyroid, but with the exception of vasomotor responses, this influence seems to be exerted indirectly through the hypophysis.

The thyroid follicle or vesicle is the anatomical and functional unit of the gland; in man its diameter is approximately 300 μ. The follicles are lined by a layer of cuboidal cells, of more or less 15-μ height, which secrete and deposit in the cavity of the follicle a colloid substance. This substance is more fluid and more intensely basophilic in the more active glands. The epithelial cells form colloid and secrete it into the acini, where it is stored. Later they reabsorb it, and there is apparently a two-way current. When the cells reabsorb colloid they emit pseudopodia into the follicle, and colloid droplets are formed in the protoplasm (Langen-

dorff, Severinghaus, De Robertis). Reabsorption seems to be carried out by a proteolytic and mucolytic process.[1] The principal structural signs of hypofunction and hyperfunction of the thyroid are summarized in Table 76.

Table 76. Structural Variations of the Thyroid in Hypofunction and Hyperfunction

	Hypofunction	*Hyperfunction*
Epithelium........	Flat	High
Follicles...........	Usually large	Size varies, many are small
Colloid............	Dense, no vacuoles* acidophil	Fluid, vacuoles* basophil
Secretory granules...	None	Many
Mitochondria......	Diminished	Increased
Golgi apparatus....	Undeveloped	Highly developed
Oxidase and cytochromoxidase granules........	....	Increased

* By means of examination *in vivo*, and the freezing and drying technique, it has been shown the vacuoles in the colloid are technical artefacts; they signiry that the colloid is fluid and retracts when fixed.

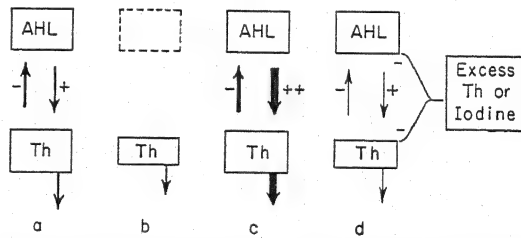
REGULATION OF THYROID FUNCTION

Interaction between the thyroid and the hypophysis. There is a reciprocal correlation between the thyroid and the hypophysis (see Chap. 52). The pars distalis of the hypophysis secretes thyrotrophin, which is a factor in the development and maintenance of the structural and functional activity of the thyroid. Removal of the hypophysis, or of the pars distalis, is followed by marked atrophy and decrease in functional activity of the thyroid. On the other hand an excess of thyrotrophin provokes hypertrophy and increases its functional activity. Treatment with thyrotrophin preparations, after a few days, provokes the appearance in the blood of an antithyrotrophic activity which neutralizes not only the thyrotrophin injected but also thyrotrophin secreted by the subject's hypophysis, thus producing thyroid atrophy and hypofunction. This rapid formation of antihormone prevents the prolonged use of thyrotrophin in therapeutics.

Thyroid secretion controls the function of the hypophysis. An increase in thyroid secretion causes a decrease in the secretion of thyrotrophin; on the contrary in hypothyroidism, or after thyroidectomy, there is an increase in thyrotrophin found in the blood and urine. Apparently there is an optimum rate of thyroid secretion which regulates normal hypophyseal function.

Thyrotrophin is selectively taken up and inactivated by thyroid and lymphoid (thymus, lymph nodes) tissue. It stimulates the thyroid, provoking the following changes:

1. *Morphologic:* formation of intracellular colloid droplets (De Robertis), increase in the height of the epithelial cells, reabsorption of colloid, increase in vascularization and weight of the thyroid (cellular hypertrophy and hyperplasia).
2. *Chemical:* the taking up of inorganic iodine by the thyroid and its conversion into organic compounds of iodine (di-iodotyrosine and thyroxine) is stimulated; organic iodine compounds (mainly thyroxine) increase.
3. *Functional:* increase in BMR and signs of hyperthyroidism (see page 601).

Circulating thyroxine inhibits thyrotrophin secretion partially or completely. On the contrary the suppression (by thyroidectomy) or a decrease of circulating thyroid hormone diminishes thyrotrophin secretion by the hypophysis (see "Antithyroid mechanisms").

Thyroid and hypophysis. *a*, normal interrelation; *b*, hypophysectomy causing atrophy of the thyroid; *c*, excess thyrotrophin secretion producing thyroid hypertrophy and hyperfunction; *d*, excess thyroid secretion or iodine causing a decrease in thyrotrophin secretion, therefore thyroid atrophy and insufficiency (there is also a direct inhibitory effect on the thyroid). The width of the arrows indicates the rate of secretion.

Thyrotrophin is found in small quantities in the urine of normal subjects. It increases in hypothyroidism, but it is not found in the urine of subjects with hyperthyroidism, apparently because it is taken up and inactivated by the thyroid.

[1] De Robertis, E., *Anat. Rec.*, **80**, 219, 1941; *Trans. Am. A. Study Goiter*, 1947, p. 1.

An excess of iodine interferes with the production and secretion of thyrotrophin by the hypophysis, inactivates thyrotrophin (Rawson), and makes the thyroid thyrotrophin-resistant.

The effect of thyrotrophin on the thyroid is also inhibited by thyroxin and some of the corticoadrenal steroids. On the other hand, thiouracil and other antithyroid drugs increase the effects of thyrotrophin (Rawson).

Factors that modify the thyroid. Most of the factors that modify thyroid structure or activity act on the anterior hypophysis, increasing or diminishing the secretion of thyrotrophin. Some of those factors also have a direct effect on the thyroid.

The principal factors that cause atrophy of the thyroid and diminish its activity are the following:

1. Extirpation of the anterior lobe of the hypophysis, which suppresses thyrotrophin secretion, necessary for the structural development and maintenance of thyroid function.
2. Excess iodine, which has a double effect: (*a*) it acts on the anterior hypophysis and diminishes thyrotrophin secretion (Loeser); (*b*) it acts directly on the thyroid, even after hypophysectomy (Chapman). Moreover, it neutralizes thyrotrophin *in vitro.*
3. Excess thyroid or thyroxine, which also has a double effect: (*a*) it decreases thyrotrophin secretion (Reforzo); (*b*) it has a direct inhibitory effect on the thyroid.[1]
4. Heat, which diminishes the activity of the thyroid.
5. Corticoadrenal hormones (cortisone, etc.) in large quantities, which act on the thyroid, diminishing the fixation of iodine and thyroid hormone secretion.
6. Several factors which, after an initial phase of stimulation, cause thyroid atrophy and insufficiency owing to exhaustion of the gland.

Many factors produce hypertrophy and hyperfunction of the thyroid by increasing the secretion of thyrotrophin or sensitizing the thyroid toward thyrotrophin. Hypertrophy, however, is not always coincident with hyperfunction; sometimes it is observed in hypofunction of the thyroid. The following are some of the principal factors that cause hypertrophy of the thyroid:

[1] Iodothyroglobulin diminishes the oxygen consumption of a surviving isolated thyroid (Galli-Mainini).

1. Excess function of the anterior hypophysis produces hypertrophy of the thyroid (increase in size), hyperplasia (increase in number of cells), reabsorption of colloid, and marked hyperactivity.
2. Compensatory hypertrophy of the thyroid is produced by extirpation of three-fourths or more of the thyroid gland; the remaining tissue undergoes a process of hypertrophy and hyperplasia. Compensatory hypertrophy does not take place if the anterior hypophysis has been removed.[1]
3. Iodine deficiency, *i.e.*, ingestion of an insufficient amount of iodine, causes hypertrophy of the thyroid, which can be prevented or cured by iodine treatment. This will be discussed in the next section.
4. A number of dietary factors tend to produce hypertrophy: (*a*) ingestion of certain proteins, such as liver protein (dogs and cats); (*b*) ingestion of certain fats, *e.g.*, butter (dogs· the hypertrophy can be prevented by cod-liver oil); (*c*) excess calcium in the diet, as occurs in rats with rickets due to an unbalanced P:Ca ratio in the diet; (*d*) deficiency in vitamin C; (*e*) preferential feeding with certain plants[2] such as some of the cabbages, rutabaga, cauliflowers, turnips, soja, etc., which provokes goiter. A crystallized antithyroid substance (L-5-vynil-2-thio-oxazolidone) has been obtained from the seed of turnips and rutabaga (Astwood, Graer, Ettlinger, 1949).
5. Cold stimulates thyroid development and activity (Cramer, Dempsey, Astwood).
6. Infections stimulate the thyroid during an initial phase; later they cause atrophy and depression. This type of response is seen in typhoid fever, tuberculosis, rheumatic fever, sepsis, etc.
7. Thiourea and thiouracil (especially propylthiouracil), thiocyanate, and some of the sulfonamides provoke thyroid hyperplasia with a marked decrease in activity (Mackenzie, Rawson).

IODINE AND THYROID FUNCTION

Iodine content of the thyroid. Baumann discovered the existence of iodine in the thyroid

[1] Houssay, Biasotti, and Magdalena demonstrated this fact in the dog and Magdalena in the toad.

[2] The most potent vegetable families are the Chenopodiaceae, Compositae, Cruciferae, Cupuliferae, Juglandaceae, Leguminosae, Rosaceae and Umbelliferae.

in 1895. He prepared an organic substance containing iodine, called iodothyrin, which was not chemically defined. Later Oswald obtained an iodized globulin (thyroglobulin). These pioneer discoveries showed the importance of iodine in thyroid function and stimulated work in this direction.

A man weighing 70 kg. has approximately 50 mg. of iodine in the whole body; from one-fifth to one-third of this amount (*i.e.*, 10 to 15 mg.) is found in the thyroid. Thyroid tissue contains 100 to 200 mg. (average 182 mg.) per 100 gm. dry weight; no other tissue has such a high concentration. Muscles have only 0.03 mg. iodine per 100 gm., but owing to the large mass of muscular tissue, approximately half the iodine in the body is found in muscles. Most of the iodine in the thyroid is in the colloid substance of the follicles; there is very little in the cells. The concentration of iodine in the colloid, or in thyroglobulin, varies considerably. The iodine content of the thyroid cannot be taken as an index of thyroid activity or thyroid secretion; it merely indicates the amount of iodine deposited in the thyroid at the time. Thus in patients with hyperthyroidism (Graves' disease) there is little iodine deposited in the thyroid, but considerable amounts are secreted, so the iodine concentration increases in the blood and tissues.

Iodine compounds in the thyroid. Iodine forms part of protein molecules, the most important of which is Oswald's thyroglobulin. Kendall (1914) obtained thyroxine from the thyroid by a process of hydrolysis; it has a very strong thyroid activity. Thyroxine has been obtained in pure crystallized form; it is a cyclic amino acid with four atoms of iodine in its molecule (iodine makes up 65 per cent of the total weight). Harington (1924–1926) established its molecular structure and later prepared it by synthesis (Harington and Barger, 1927). Di-iodotyrosine (Harington and Randall, 1929) and mono-iodotyrosine (Fink, 1948) have been isolated from thyroid tissue. Another iodine compound, 3:5:3-L-tri-iodothyronine, is more potent and acts more rapidly than thyroxine in many of its effects. It has been suggested that thyroxine is converted into tri-iodothyronine in the cells where it is active. Tri-iodothyronine, therefore, would be the active form of thyroid hormone in the periphery.[1]

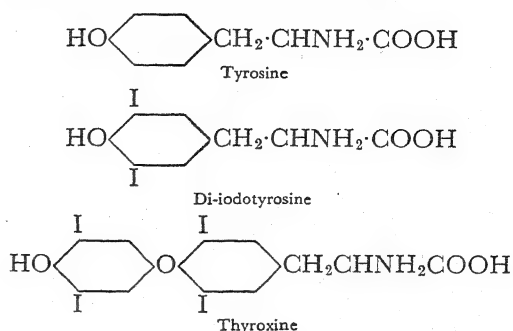
Thyroxine ($C_{15}H_{11}O_4NI_4$), also called tetra-iodothyronine,[1] is L-β[3,5-di-iodo-4(3′,5′-di-iodo-4′-hydroxyphenoxy)-phenyl] -α-amino-propionic acid. Possibly it is formed from tyrosine, which is first converted into di-iodotyrosine, then into thyroxine.

HO–⟨ring⟩–$CH_2 \cdot CHNH_2 \cdot COOH$

Tyrosine

HO–⟨ring, I, I⟩–$CH_2 \cdot CHNH_2 \cdot COOH$

Di-iodotyrosine

HO–⟨ring, I, I⟩–O–⟨ring, I, I⟩–CH_2CHNH_2COOH

Thyroxine

The total amount of thyroxine in the body is calculated to be from 8 to 14 mg.; and the daily consumption of thyroxine 0.25 mg. (Boothby) or 0.33 mg. (Plummer).

Thyroxine produces a marked increase in the BMR after a latent period of 7 to 48 hr. Its effect lasts 1 to 10 weeks (Fig. 245). This remarkable activity is similar to that of enzymes. One dose of 1 mg. of thyroxine can provoke a 2.5 per cent increase in the BMR and a total production of 1,000 extra calories. Adrenalectomy causes a considerable decrease in the calorigenic effect of the thyroxine in the rat; the administration of corticoadrenal hormones restores the capacity of the animals to respond to thyroxine.[2]

The activity of L-thyroxine is much greater than that of D-thyroxine. The substitution of Br for I reduces activity to one-seventeenth, and that of Cl for I reduces it to one two-hundred-and-fiftieth.

Iodine in the form of iodide is rapidly taken up by the thyroid and converted into ionic or active iodine, apparently by oxidation by an enzymatic system, of which possibly iodine itself forms part. In a further stage the thyroid hormone is formed, by iodization of tyrosine to mono-iodotyrosine then di-iodotyrosine, and oxidative coupling of two molecules of the latter into thyroxine, which is stored as thyroglobulin in the colloid and finally secreted into the blood.

[1] Gross, J., and R. Pitt-Rivers, *Lancet*, **262**, 593, 1952.

[1] Thyronine has the same structure as thyroxine, but the four iodine atoms are replaced by four atoms of hydrogen.

[2] Hoffmann, F., *et al.*, *J. Physiol.*, **107**, 251, 1948; *Acta physiol. Latinoam.*, **1**, 84, 1951.

Formation of thyroxine outside the thyroid. Thyroxine can be produced outside the thyroid. By treating casein and other proteins with iodine *in vitro*, thyroxine is produced (Ludwig and Mutzenbecher, 1939). If radioactive iodine is given to thyroidectomized animals, a small amount of radioactive thyroxine is formed. Small quantities of thyroxine are ingested with meat (muscle). Iodized proteins have a similar, but not as potent, activity as thyroid hormone. They are active by mouth and have been used with the object of increasing milk secretion or the production of eggs.

Blood iodine. The iodine content of the blood is not easily measured; values found vary between 5 and 20 μg per 100 cc., according to the method used. Iodine that precipitates together with plasma or serum protein—protein-bound iodine (PBI) or hormonal or precipitable iodine—is of special interest in medical practice, because it is a good index of thyroid function in subjects which do not receive iodine in the food, or iodine treatment. More than 90 per cent PBI is in the form of thyroxine[1] mainly in serum albumin and in smaller amounts in α and β globulins and other proteins. Triiodothyronine, which is more active than thyroxine, has also been found. Thyroxine seems to be the preferential form of secretion, since it makes up 90 per cent of blood iodine, and only 26 to 30 per cent of thyroid iodine.

Protein-bound iodine is found in a concentration of 4 to 8 μg (average 5.2 μg) per 100 cc. of plasma or serum in subjects with normal thyroid function; it increases above 8 μg per 100 cc. in hyperthyroidism and falls to 3 μg per 100 cc. or less in hypothyroidism. Stimulation of the thyroid by thyrotrophin provokes a rise in PBI. It increases slightly in pregnancy (6 to 11 μg). Thyroidectomy, hypophysectomy, and treatment with antithyroid drugs cause a decrease in PBI.

A homeostatic mechanism regulates synthesis of organic iodized substances in the thyroid.[2] If the concentration of iodine in blood rises above 15 to 25 μg per 100 cc., iodine is no longer bound to organic substances in the thyroid; when blood-iodine concentration falls, iodine is again bound to organic substances in the thyroid.

[1] TAUROGG, A., *et al.*, *J. Biol. Chem.*, **184**, 99, 1950; GORDON, A. H., *et al.*, *Nature, London*, **169**, 19, 1952.

[2] WOLFF, J., and I. L. CHAIKOFF, *Endocrinology*, **42**, 468, 1948; *J. Biol. Chem.*, **174**, 555, 1948.

Fixation and utilization of iodine by the thyroid. Radioactive isotopes of iodine[1] have been important aids for the advancement of knowledge of thyroid function. They have been used in the study of (*a*) the uptake of iodine by the thyroid (amount, and time-concentration curve); (*b*) its combinations with organic substances in the thyroid, especially its appearance in thyroxine; (*c*) thyroid secretion, indicated by the speed at which it leaves the thyroid and increases in protein-bound iodine in the blood; (*d*) its elimination in the urine.

Iodine is essential for the production and control of the secretion of thyroid hormone. It is obtained from foodstuffs. In the course of the catabolism of thyroid hormone, part of its iodine is excreted in the urine, and part is retained and again used by the organism.

The amount of iodine taken up by the thyroid depends on the condition of the gland and the secretion of thyrotrophin. Thyrotrophin increases the uptake of iodine, its incorporation into organic compounds (thyroxine), and its secretion into the blood. It also provokes hypertrophy, hyperplasia, and hyperfunction of the thyroid.

Deficiency of thyrotrophin, due to hypophyseal insufficiency, causes atrophy of the thyroid, which weighs less and has a low, hypofunctioning follicular epithelium. The uptake of iodine is diminished, but as there is a low rate of secretion, the concentration of iodine in the gland remains normal. The protein-bound iodine in blood is below normal.

A hyperplastic thyroid[2] takes up more iodine than a normal gland and converts it more rapidly into thyroxine. This activity is observed in exophthalmic goiter, in which there are thyroid hyperplasia and hyperfunction, and in endemic goiter, in which there is hyperplasia without hyperfunction. Antithyroid drugs inhibit the uptake of radioactive iodine by hyperplastic thyroid tissue.

In hypothyroidism (*e.g.*, myxedema) and after hypophysectomy the uptake of radioactive iodine is diminished. Antithyroid drugs also diminish the uptake of iodine, but when the

[1] Isotopes of iodine are I^{130}, I^{131}, I^{125}, and astatine (element 85). I^{131} is the one most frequently used in physiology and therapeutics.

[2] In nodular hyperplasia the nodules can be located by giving the patient radioactive iodine and applying a radiation detector (Geiger-Müller counter).

treatment is discontinued, if it has provoked thyroid hyperplasia, the uptake is above normal. The thyroid and other tissues of subjects with hypothyroidism given radioactive iodine take up less iodine, and in 24 to 72 hr. excrete more in the urine, than normal subjects. In hyperthyroid subjects the reverse occurs, *i.e.*, they take up more iodine than normal subjects and excrete less.

If large quantities of radioactive iodine are incorporated into the thyroid, the β radiation may destroy part or even the whole gland. This effect has been used in the treatment of hyperthyroidism and, in larger doses, in the treatment of cancer of the thyroid. The results vary according to the capacity of the neoplastic tissue for taking up radioactive iodine.

Response of the thyroid to iodine deficiency or excess. When the iodine content of the thyroid falls below 100 mg. per 100 gm. of tissue, the thyroid becomes hypertrophied (Marine). The process occurs in the following stages: (*a*) the height of the cells lining the follicles increases and the thyroid increases in size (hypertrophy); (*b*) the cells multiply and increase in number (hyperplasia); (*c*) the gland is exhausted and becomes atrophied or large amounts of colloid are accumulated in the follicles (colloid goiter). A gland in this condition of lower activity or at rest can again become active and pass into hypertrophy, hyperplasia, etc., repeating the cycle several times.

Normal → hypertrophy → hyperplasia → atrophy due to exhaustion
↘ colloid
↓
hypertrophy → hyperplasia → atrophy due to exhaustion
↘ colloid

Hypertrophy and hyperplasia of the thyroid can be produced experimentally by feeding animals on diets deficient in iodine. Thyroid response to iodine deficiency is provoked by a double mechanism; (*a*) there is an increase in thyrotrophin secretion; (*b*) there is a direct effect on the thyroid. The direct effect has been demonstrated by feeding hypophysectomized rats a diet deficient in iodine. In these circumstances the thyroid responded with hypertrophy and hyperplasia, but less markedly than the controls (Chapman). Thyroid hyperplasia can be prevented or cured by the administration of iodine (Marine); the cells become flattened, and the follicles are filled with colloid. This inhibitory effect of iodine is observed whenever factors, such as thyrotrophin, etc., which cause thyroid hypertrophy, become active. It is utilized in therapeutics in order to control excessive thyroid function, especially in thyrotoxicosis. Administration of iodine provokes a decrease in the BMR and improves the general condition of such patients, so that they can be submitted to subtotal thyroidectomy with considerably less risk. Iodine treatment prevents the appearance of the so-called "postoperative crises" and has greatly diminished operative mortality. If iodine treatment is unduly prolonged, symptoms of hyperthyroidism reappear with the initial or even greater severity. In these cases it is said that the patient has become refractory or resistant to iodine; therefore iodine treatment should not be administered to patients suffering from hyperthyroidism except under strict medical supervision. Propylthiouracil and mercaptoimidazol are now used in therapeutics instead of or together with iodine, in order to control excess thyroid activity.

Iodine has a double effect on the hypophysothyroid mechanism: (*a*) it diminishes the secretion of thyrotrophin by the hypophysis (Loeser), and (*b*) it inactivates thyrotrophin (Rawson). The inhibitory effect on the thyroid is due to (*a*) inhibition of enzyme systems, and (*b*) inhibition of the proteolytic enzyme which takes part in the reabsorption of colloid (De Robertis). Small doses of iodine stimulate the thyroid; medium and large doses inhibit it, but after a time they lose their effect. Hyperplastic thyroids are more sensitive to the inhibitory action of iodine than normal thyroids.

Prophylaxis of goiter and cretinism. Endemic goiter and cretinism are due to iodine deficiency; therefore iodine administration is the

fundamental factor in the prophylaxis of these diseases. Chatin in 1850 and later Fellenberg[1] showed that in districts where goiter was endemic the iodine content of the air, soil, water, and foods, and the daily ingestion of iodine, were lower than in districts free from goiter. The thyroid is enlarged and shows signs of hyperplasia in a large percentage of the population of districts deficient in iodine. Boussingault in 1831 and later Marine (1907–1911) and many others have demonstrated that the administration of adequate doses of iodine not only has a preventive effect on goiter, but also diminishes the incidence of endemic goiter, cretinism, deaf-mutism, and thyrotoxicosis and improves the general condition (*e.g.*, increases the average height) of the population in goitrous areas. The existence in water of a factor causing goiter has been postulated, among other reasons because thyroid hyperplasia can be provoked in animals reared in nongoitrous areas by giving them water from a goitrous area (Bircher, Houssay). This experimental thyroid hyperplasia can also be prevented or cured by iodine.

The incidence of goiter increases toward the end of winter. It is more frequent in women than in men, especially in adolescents and in pregnant women. Iodine treatment should be given during pregnancy, in goitrous districts, in order to assure the healthy development of the fetus. Cretins are born in goitrous areas from mothers with goiter, if they have not been given iodine.

The most efficient method of goiter control consists in the addition of iodine to table salt (usually KI and a stabilizer to prevent loss of iodine). In Switzerland 5 mg. iodide is added to each kilogram of salt. In England the addition of 10 mg. per kg. of common salt and 25 mg. per kg. of table salt has been advised.[2] Larger amounts of iodide (100 mg. per kg. of salt) have been considered advisable for goitrous areas in the United States,[3] and 10 mg. per kg. (Kimball) for the whole country. The optimum amount of iodine for a normal subject is 100 to 200 μg daily; two or three times this dose should be administered in goitrous areas.

[1] Fellenberg, T., *Ergebn. d. Physiol.*, **25**, 176, 1926.

[2] Salt consumption per capita has been calculated to be 10 gm. per day; therefore 100 μg iodide, equivalent to 76 μg iodine, would be the average amount consumed (*Lancet*, **1**, 107, 1944; **1**, 913, 1948).

[3] The average consumption of salt is calculated to be 6.2 gm. per day, therefore 620 μg iodide, equivalent to 470 μg iodine, is given in this case. (Curtiss, G. M., and M. B. Fertman, *J. A. M. A.*, **121**, 423, 1943. See *J. A. M. A.*, **130**, 81, 1946; **135**, 434, 1947.)

Antithyroid mechanisms. The thyroid gland can be inhibited by several mechanisms:

1. The secretion of thyrotrophin is diminished by the effect of thyroid, thyroxine, or iodine on the hypophysis and by irradiation of the hypophysis.
2. Thyrotrophin is inactivated by iodine (Rawson) and by antithyroid serum prepared by immunization (Harington).
3. Iodine uptake by the thyroid is inhibited by thiocyanate, which also releases iodine from the thyroid. As there is no iodine in the thyroid, organic iodine compounds including thyroxine are not synthesized. The normal inhibitory effect of thyroxine on the hypophysis is absent; therefore thyrotrophin secretion increases and provokes thyroid hyperplasia and goiter. The administration of iodine suppresses all these effects. When treatment with thiocyanate is suspended, the hyperplastic thyroid takes up iodine at a high rate.
4. Thiourea and thiouracil derivatives inhibit the production of thyroid hormone. The uptake of iodine may or may not be below normal, but active iodine is not formed and bound in organic compounds, nor is thyroxine synthesized. Thus the inhibitory action of thyroxine on the hypophysis diminishes, and thyrotrophin secretion increases, causing thyroid hyperplasia. Iodine is not very efficacious in counteracting the effect of these drugs. Methyl-thiouracil, propyl-thiouracil and later 1-methyl-2-mercaptothiouracil have been used in the treatment of hyperthyroidism; they are employed before thyroidectomy in order to diminish the size and activity of the thyroid, thus facilitating the operation.
5. Agents that inhibit the effect of thyroid hormone on the tissues are not well known.

THE FUNCTIONS OF THE THYROID

Methods of study. The principal methods for the study of thyroid function are:

1. The anatomical method, consisting in the observation of gross and microscopic changes in the thyroid under different conditions.
2. Observation of spontaneous or experimental thyroid insufficiency. Experimental thyroid insufficiency can be provoked by (*a*) thyroidectomy;[1]

[1] Early experimenters unwittingly removed the para-

(*b*) irradiation of the thyroid with x-rays or radioactive iodine; (*c*) hypophysectomy; (*d*) antithyrotrophic agents; (*e*) inhibition of thyroid secretion (iodine, thiouracil, etc.). Knowledge of thyroid function in man was first obtained by the observation of spontaneous myxedema,[1] postoperative myxedema,[2] and cretinism. Symptoms observed should be attributed to thyroid insufficiency only if they are controlled by "substitution therapy," *i.e.*, by restoring thyroid function by thyroid grafts or by the administration of thyroid by mouth or by injections of thyroxine.[3]

3. Observation of spontaneous or experimental hyperthyroidism. Experimental hyperthyroidism can be provoked by the administration of an excess of thyroid, or injections of thyroxine or thyroglobulin, or injections of thyrotrophin if the thyroid is intact and capable of responding. Pathologic or spontaneous hyperthyroidism is observed most frequently in man in cases of Graves' disease or toxic adenoma of the thyroid.
4. Tyroid function can be explored by (*a*) determination of the BMR; (*b*) determination of protein-bound iodine in the blood; (*c*) administration of radioactive iodine in order to measure its uptake by, and concentration in, the thyroid and the rate of its elimination in the urine.

Morphogenetic Function of the Thyroid in Amphibians

Thyroid secretion is a necessary factor for metamorphosis in amphibians; *i.e.*, for reabsorption of the tail and gills and the development of the limbs. Thyroidectomized tadpoles do not develop for lack of the thyroid (Allen); hypophysectomized tadpoles do not develop because the thyroid atrophies (Smith). The axolotl remains in the larval state because it has a rudimentary thyroid. The administration of thyroid, on the contrary, accelerates metamorphosis in normal and thyroidectomized or hypophysectomized tadpoles. If thyroid treatment is given prematurely, metamorphosis takes place before the animals have grown to their normal size, and the resulting adult frogs are very small (Gudernatsch). The results of thyroid treatment are particularly remarkable in the axolotl. This species remains all its life in an aquatic larval condition, with external gills, reproducing itself without metamorphosis into an adult form. Thyroid treatment converts it into a terrestrial salamander. Acceleration of metamorphosis is due to iodine and can be brought about not only by thyroid, thyroxine, or di-iodotyrosine, but also by iodine alone (Swingle).

thyroids at the same time as the thyroid, and attributed many of the effects of parathyroid insufficiency to the thyroid. Schiff was the first to perform thyroidectomy (1858).

[1] Gull, 1873; Ord, 1877.

[2] Reverdin, 1879–1882; Kocher, 1883, who gave the name "cachexia strumipriva" to the condition.

[3] Fresh or powdered thyroid can be given by mouth because the active principle resists digestion. Thyroxine is given by injections because it is partially inactivated when given by mouth.

a

b

FIG. 243. Effect of thyroidectomy on growth. *a*, thyroidectomized horse on the right, control on the left; *b*, thyroidectomized goat on the left, control on the right.

Thyroid Insufficiency

Growth and morphogenesis. The earlier the age at which thyroid insufficiency is established, the greater the severity of its effects. They are especially severe when the disturbance is initiated during fetal life[1] or during the first months after birth. Insufficiency is more pronounced if the iodine in the diet is inadequate.[2] The sever-

[1] Hence the importance of iodine treatment in pregnant women living in a goitrous district.

[2] LEBLOND, C. P., and H. EARTHY, *Endocrinology*, 51, 26, 1952.

ity also varies with the species; man is one of the most sensitive and suffers most from thyroid deficiency.

Thyroidectomy in the young of all vertebrate species arrests growth and development; the animals do not attain a normal size and do not develop harmoniously. Animals in which thyroid deficiency begins during fetal life or in the first years after birth are cretins and dwarfs (Figs. 243 and 244). Their height is subnormal. Ossification is markedly retarded; the bones of the face grow less than those of the cranium, the nose is flattened, and the limbs are short. The abdomen is prominent, with a protruding umbilicus. The eruption of the permanent teeth is retarded, and they are frequently irregularly placed, badly developed, and poorly calcified; there is a high incidence of caries. In man the skin is dry, rough, and cold. There is a characteristic infiltration of the subcutaneous tissues, which is soft but does not "pit" on pressure. This infiltration was called "myxedema" by Ord; it is marked in man but is not seen in other species. The face is round and puffed out, with swollen eyelids (Fig. 246); the skin is usually pale, but sometimes the cheeks are red. The tongue is swollen and sometimes protrudes between thickened lips. The hair is coarse and straight; frequently it falls in patches or diffusely. The eyebrows are thin, and there is little hair in the axillae. Thyroid treatment, especially if it is begun at an early age, cures all these disturbances, but if it is delayed the results are less satisfactory.

Fig. 244. Congenital cretinism. Age, twenty-seven years; weight, 41 kg.; height, 1.19 m.; BMR, −19 per cent. (*Courtesy of Dr. E. B. del Castillo.*)

Basal metabolism. Thyroidectomy is followed by a marked decrease in heat production (Fig. 245). The BMR falls 30 to 45 per cent below the normal.[1] The fall takes place gradually, reaching its lowest level in 60 to 80 days after thyroidectomy. Myxedema appears later. The skin is cold, and the body temperature is frequently below normal. Thyroidectomized subjects are usually very sensitive to cold and they do not tolerate low temperature of 0 to 2°C. as well as normal subjects. In conditions causing fever or hyperthermia, the rise in temperature is less than in subjects with normal thyroid function. The specific dynamic action of food is low, partly because of slow intestinal absorption. Thyroidectomized animals are more resistant to anoxia than normal animals.[2]

Thyroid preparations given by mouth or injections of thyroxine restore the BMR to a normal or even higher level in cases of thyroid insufficiency and in thyroidectomized animals (Fig. 245).

[1] It is important to keep in mind that the majority of subjects with a BMR of −15 to −20 per cent do not suffer from thyroid insufficiency (Boothby).

[2] Duran, Streuli, De Quervain, Rietti, Leblond, etc.

Protein metabolism. Animals with thyroid insufficiency take up less nitrogen for the formation of body protein than normal animals. This is especially marked in young animals, in which growth ceases. Protein catabolism is also subnormal, as is shown by the low nitrogen excretion in total or protein fasting.

Fat metabolism. There is usually a high lipemia in cases of thyroid insufficiency, and plasma cholesterol is increased. The fat content of the liver also increases. Occasionally there is accumulation of storage fat, but hypothyroidism is not the cause of most cases of obesity. In severe thyroid insufficiency there is loss of appetite, the subjects lose fat and weight, and finally end in cachexia.[1]

Carbohydrate metabolism. Blood sugar and liver and muscle glycogen are usually normal or subnormal, except in cases with cachexia, in which there is a definite decrease. The intestinal absorption of glucose and galactose is retarded (Althausen), and the blood-sugar curves following ingestion of these sugars are flattened and prolonged compared with those observed in normal subjects. Sugar is normally oxidized by the tissues. Sensitiveness to insulin is usually increased in thyroidectomized animals; insulin hypoglycemia is more marked and prolonged, and convulsions occur with greater frequency than in normal animals.[2]

Mineral and water metabolism. The passage of water through the body is usually retarded in thyroid insufficiency; thus after ingestion of water diuresis appears later than in normal subjects and is less marked but more prolonged. Myxedema (Fig. 246) is caused by the accumulation of water, sodium chloride, and over 13 per cent of protein. This protein has the appearance of mucin, the protein in mucus (whence the name "myxedema"), but its chemical nature is not well known. Young individuals retain less calcium and phosphorus than normal ones, owing to retarded ossification and underdevelopment of the skeleton, but the bone formed is usually compact and resistant. Protein-bound iodine in blood is diminished. If radioactive iodine is given, less than normal amounts are taken up by the thyroid and more is excreted in the urine. In a previous paragraph the importance of the thyroid as the main factor in iodine metabolism was fully discussed.

[1] This severe form was observed after total thyroidectomy performed for the treatment of goiter (Kocher's "cachexia strumipriva"); now usually subtotal thyroidectomy is practiced.

[2] Ducheneau, 1923; Bodansky, 1923; Houssay and Busso, 1924.

Blood and circulation. Secondary anemia is frequently observed; it is usually slight or moderate; in exceptional cases it may be of considerable severity. There is a typical increase in plasma globulins and in the viscosity of the blood.[1]

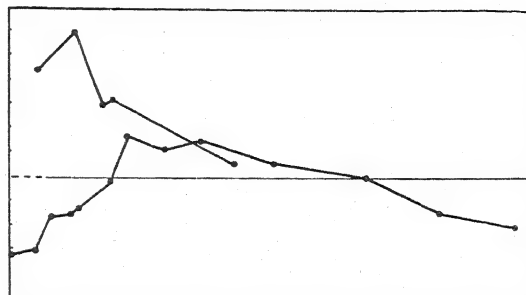



FIG. 245. Effect of thyroidectomy and thyroxine on the BMR. Upper curve, BMR of patient with adenoma of the thyroid with hyperthyroidism, before and after thyroidectomy; lower curve, BMR of patient with myxedema, effect of two intravenous injections of 8 and 16 mg. of thyroxin. (*Boothby, W. M., and Irene Sandiford, Physiol. Rev., vol. 4, p. 104, 1924.*)

Circulating plasma volume and minute volume decrease in thyroid insufficiency; the latter may fall to 2.5 liters per min. The circulation is slow and the circulation time is therefore increased.[2] Intracellular and extracellular edema of the heart is observed in some cases of thyroid insufficiency, which may lead to cardiac dilatation and a weak heartbeat. The deflections of the electrocardiogram are of subnormal amplitude, and the T wave is sometimes inverted.

Changes in the blood and circulation usually respond satisfactorily to thyroid treatment.

Digestive functions. There is usually a poor appetite and constipation is frequently observed.

Nervous and psychic disturbances. These are particularly marked in human cases of thyroid insufficiency, especially those which occur at an early age, because the thyroid is an important factor in the normal development of the central nervous system. The patients are apa-

[1] Fano and Rossi, 1905; Rossignoli *et al.*, 1932; Goldberg, 1938.

[2] Thyroidectomy has been performed in some cases of irreducible circulatory insufficiency and of severe angina pectoris, in order to slow down the circulation and diminish the load on the heart.

thetic, their mental processes are slow, they have a poor memory, and they are frequently obtuse. Children do not learn easily and are backward in their school rating.

The muscles are frequently weak and easily fatigued. On percussion, the muscles sometimes respond with a localized contracture. Muscle tonus is usually decreased, causing laxitude of the joints and abdominal dilatation.[1] Speech is slow and the voice has a peculiar intonation owing to myxedematous infiltration of the laryngeal mucosa.

Immunity. The skin and mucosae are less resistant to infections in hypothyroid individuals than in normal ones. The phagocytic capacity of the leukocytes is subnormal. The formation of hemolysins and agglutinins is usually normal, but the production of antitoxin is subnormal in thyroidectomized animals. It is difficult to sensitize these animals, but if they have been sensitized before thyroidectomy, anaphylactic shock is easily provoked.[2]

[1] Hypothyroid subjects with myotonia have muscular hypertrophy (Lanari).

[2] HOUSSAY, B. A., and A. SORDELLI, "Tiroides e Inmunidad," Espasa-Calpe, Madrid, 1924.

Thymus. The thymus gland usually retrogresses at an early age in thyroidectomized animals, but there are no constant changes in this organ.

Sexual functions. Changes observed in thyroid insufficiency will be considered in Sec. VII.

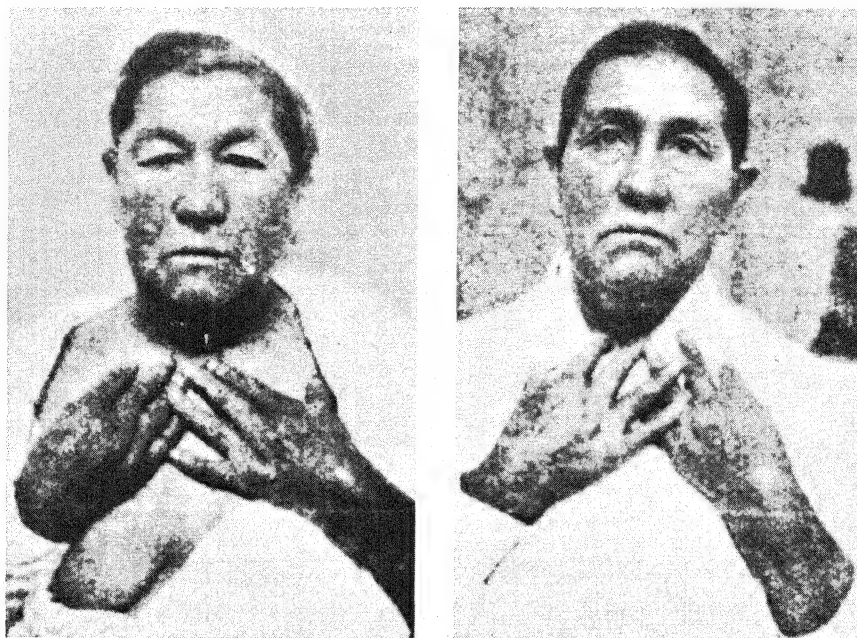



FIG. 246. Myxedema in woman forty-nine years old. Left, before treatment, marked infiltration on hands, face, and eyelids, BMR −35 per cent; right, after 1½ months' thyroid treatment, BMR +5 per cent. (*Courtesy of Dr. E. B. del Castillo.*)

Treatment. Thyroid insufficiency is controlled by the so-called "substitution treatment." There are three methods: (*a*) thyroid graft; (*b*) oral administration of thyroid preparation; (*c*) injection or oral administration of thyroxine. Thyroid grafts have been made in animals and man; if they "take," *i.e.*, if vascular connections are established and the grafted tissue survives, all the effects of thyroid insufficiency are cured. Hyperthyroidism cannot be provoked by multiple grafts; therefore it is advisable to perform many small grafts in order to assure the survival of an adequate amount of glandular tissue.

Oral administration of thyroid preparations in a case of myxedema was first practiced by Murray in 1891, and the patient survived in good health for many years. This method is the one most commonly used for the treatment of thyroid insufficiency, because the active princi-

ple of the thyroid resists the effects of digestion and is easily absorbed. Furthermore, the dose can be accurately controlled and the treatment suspended whenever it is advisable.

Thyroxine (the sodium salt) is given in intravenous injection. It has a prolonged effect, so that the injections can be given at fairly long intervals, but if an excessively large dose has been administered or the patient is abnormally sensitive, the undesired effects last for several days.

Small doses of thyroid preparations stimulate the anterior hypophysis in some cases, but large doses depress this gland (see Chap. 52).

Hyperthyroidism

Excessive thyroid function provokes disturbances, which are particularly marked in man. In the following description, therefore, the symptoms of human hyperthyroidism will be emphasized.

Basal metabolism. Oxygen consumption and heat production increase in hyperthyroidism. The increase in the BMR is a sign of considerable importance for the diagnosis and treatment of hyperthyroidism, because it is an accurate measure of the level of thyroid function. Oral administration of thyroid preparations and injection of thyroxine also provoke an increase in the BMR. The body temperature is usually high but within the normal range; exceptionally it may rise a little above the normal. The skin of the hands and face is frequently warm or hot, but this is not a sign of fever.

Anoxia is not well tolerated by animals with hyperthyroidism, probably because of the greater need of oxygen caused by the increase in heat production. If they are submitted to low oxygen tension they die at higher oxygen partial pressures than the controls. This contrasts with the greater resistance to anoxia observed in thyroid insufficiency.

Protein metabolism. Protein catabolism increases in hyperthyroidism, and the minimum protein requirement is higher than in normal subjects. This fact, reported by F. von Müller in 1893, was one of the first disturbances in metabolism to be observed in hyperthyroid subjects. It led Magnus Levy (1895) to the discovery of the increase in oxygen consumption and heat production typical of the condition. Thyroid treatment also increases protein catabolism (increased urinary elimination of nitrogen) in adults. In young animals, especially in hypothyroid dwarfs, small doses of thyroid cause an increase in the retention of nitrogen owing to the formation of body protein needed for growth; but even in the young, large doses of thyroid increase nitrogen excretion. Creatinuria and a high creatinine excretion are frequently observed in hyperthyroid subjects.

Fat metabolism. A popular fallacy attributes to the thyroid the capacity of especially accelerating fat consumption, and thyroid treatment is given in order to cure obesity. The thyroid does increase the oxidation of fat, but not in preference to protein or carbohydrate.

Hyperthyroidism, especially if it is marked, rapidly leads to loss of fat. Ketosis is also frequently observed, because carbohydrate stores are depleted and the high metabolic rate has to be kept up at the expense of storage fat; the capacity of the tissues to burn fat is thus often exceeded. This should be kept in mind when treating diabetic patients suffering from hyperthyroidism; they easily fall into ketosis and acidosis. Blood cholesterol is usually low and sometimes below normal.

Carbohydrate metabolism. The fasting blood-sugar level of hyperthyroid subjects is usually normal or slightly above normal. The postabsorptive increase in blood sugar is generally greater than that observed in normal subjects; it frequently rises above the renal threshold and causes a small degree of glycosuria. Selective intestinal absorption of glucose, and especially of galactose, takes place at a higher rate than normal (Althausen) and is the cause of the rapid increase in blood sugar and the marked and prolonged hyperglycemia observed after ingestion of these sugars (Figs. 201 and 202). These high and prolonged hyperglycemic curves following the absorption of sugar may also be due partly to a lower rate of glycogen formation by the liver. These curves differ from diabetic hyperglycemic postabsorptive curves because the blood sugar returns to a normal level. Moreover during hyperglycemia the oxidation of glucose is increased in hyperthyroidism, but not in diabetes.

Thyroid treatment causes a rapid decrease in liver glycogen; the decrease is particularly rapid after a few hours of fasting. The formation of glycogen by the liver after the ingestion of sugar is also delayed in severe hyperthyroidism. Liver glycogen is the most sensitive to an increase in

thyroid function; heart glycogen decreases later; muscle glycogen is diminished only in severe and prolonged hyperthyroidism.

Insulin resistance is usually, but not always, increased by a slight or moderate increase in thyroid function. In severe hyperthyroidism, when the glycogen stores are depleted, there is frequently hypoglycemia in fasting and considerable hypersensitiveness to insulin.

Experimental or spontaneous hyperthyroidism does not provoke diabetes if the pancreas is intact, but if the pancreas has been partially extirpated or the islets damaged (*e.g.*, by anterior hypophyseal diabetogenic extract), large doses of thyroid preparations given over a sufficiently long period may provoke diabetes by further damaging the β cells of the pancreatic islets. This thyroid diabetes is transitory, and the animals return to a normal condition after the thyroid treatment is discontinued, if it has not been given for an excessively long time. On the other hand, if the thyroid treatment has been kept up for several weeks, the islet lesions become irreversible, the β cells are destroyed, and the animals remain permanently diabetic even after the thyroid treatment is discontinued (metathyroid diabetes). The incidence of diabetes in hyperthyroid subjects is twice that in the general population. This is due probably to further damage of an already weakened pancreas caused by excessive thyroid secretion (see Chap. 41).

Mineral and water metabolism. In the first stages of experimental hyperthyroidism, or during the first days of thyroid treatment, diuresis is increased, and there is loss of weight due to loss of body water.[1] Thyroid treatment in myxedematous patients causes a considerable increase in the urinary elimination of water, sodium chloride, and urea, which is due to the mobilization of the myxedematous fluid. The elimination of ingested water is performed slowly in hyperthyroid subjects; edema, especially of the eyelids and around the ankles, may be observed after copious drinking of water in these patients.

Protein-bound iodine in blood is increased. The thyroid takes up more radioactive iodine than a normal gland, and the urinary excretion of radioactive iodine in the 48 to 72 hr. after it has been given is below normal.

[1] This loss of weight is often mistakenly attributed to oxidation of fat.

During the initial stages of hyperthyroidism and in hyperthyroid crises, calcium and phosphorus elimination in the urine and feces is increased, and there is a negative Ca and P balance. After a time the calcium and phosphorus equilibriums are restored; nevertheless the bones of hyperthyroid patients are often poorly calcified and sometimes osteoporosis is observed.

Growth. In some cases of hyperthyroidism in infancy and adolescence, a slight acceleration of growth has been observed; but in the majority of cases, although the period of growth is prolonged, the normal height is not surpassed. Giantism is never observed in human or experimental hyperthyroidism, and if an excessively large dose of thyroid is given, growth may be below normal. Hyperthyroid subjects are usually thin. Some of them have prematurely gray hair. The skin is usually warm and moist.

Blood and circulation. Lymphocytosis is often seen in hyperthyroid subjects, but it also occurs in cases of thyroid insufficiency. The phagocytic capacity of the leukocytes is usually increased in hyperthyroidism. Plasma proteins are decreased, and the viscosity of the blood is subnormal.

The circulating blood volume increases in hyperthyroidism, and there are typical and marked disturbances in the heart such as tachycardia, palpitations, a large minute volume, and a rapid circulation with a reduced circulation time. Crises of vasodilatation (hot flashes) are often observed.

Nervous and psychic disturbances. Animals with experimental hyperthyroidism are restless and irritable. Patients with hyperthyroidism, especially those with exophthalmic goiter, are excessively emotional and unstable. Insomnia and general irritability and nervousness often demand rest in bed and sometimes complete quiet and seclusion. There is hypersensitiveness to adrenaline, which provokes pseudoaffective crises (tremor, anguish, weeping).

Hyperthyroid patients are easily fatigued and muscular efficiency is below the normal. Muscular weakness and slight or even marked muscular atrophy are often seen.

Hoffmann[1] has observed that in experimental hyperthyroidism peripheral nerve stimulation pro-

[1] Hoffmann, F., and E. J. de Hoffmann, *Bol. Soc. biol. de Chile*, **5**, 44, 1948.

duces less effect than in normal animals, a higher frequency of stimulation is needed to provoke tetanus, exhaustion occurs rapidly, and the recovery period is prolonged. Stimulation of the vagus nerve often fails to stop the heartbeat, and if it does stop, "vagus escape" soon occurs. The response to acetylcholine is diminished, but during peripheral nerve stimulation a substance is released which causes contraction in the denervated eye muscles and circulatory collapse.

Digestive functions. Hyperthyroid patients usually have a large appetite. Diarrhea occurs in hyperthyroid crises and in severe cases. It is often accompanied by vomiting and causes rapid loss of weight.

Thyroid intoxication. Large doses of thyroid preparations produce polyuria, loss of weight, weakness, nervousness, insomnia, tachycardia, palpitations, diarrhea, acidosis, and severe lesions in the liver, heart, and other tissues. If large doses are given over a long period they may be fatal. In the advanced stages of thyroid intoxication, muscle glycogen is decreased, the blood sugar is low, and hypoglycemic crises are easily provoked by insulin or occur after the hyperglycemic response to glucose ingestion; in the rabbit hypoglycemic crises occur spontaneously.

Vitamins. Vitamin A causes a decrease in the severity of the disturbances caused by hyperthyroidism in the rat, but in man this effect has not been clearly demonstrated. The fall in liver glycogen provoked by thyroid treatment is diminished by the administration of vitamins B_1 and B_2.

The increase in metabolism which is the consequence of hyperthyroidism apparently increases vitamin requirements. It is therefore advisable to give hyperthyroid patients larger amounts of vitamins than those considered adequate for normal individuals.

Abnormal Thyroid Conditions in Man

Goiter. This is the name given to the enlargement of the thyroid, due to an increase in thyroid tissue or to the accumulation of colloid. Goiter is compatible with a normal thyroid function, and most people with endemic goiter in goitrous districts do not suffer from a disturbed thyroid secretion. In some cases goiter is accompanied by thyroid insufficiency, in others by hyperthyroidism.

Cretinism. Thyroid insufficiency occurring at an early age causes deficient physical and mental development. The subjects may be complete idiots. They are usually of small stature with short limbs and a disturbance in the development of the bones in the face and cranium which gives them a typical aspect (Fig. 244). There may be a goiter, or the thyroid may be atrophic. Cretinism is observed in goitrous areas, but occasionally a case is seen outside these areas. Iodine administration during pregnancy and lacta-

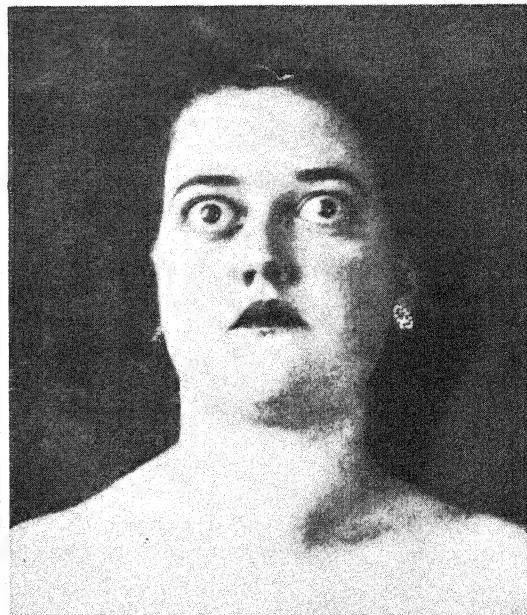



Fig. 247. Exophthalmic goiter. BMR +55 per cent. (*Courtesy of Dr. E. B. del Castillo.*)

tion prevents the appearance of cretinism in the offspring.

Myxedema. Primary myxedema may occur spontaneously, *i.e.*, without an apparent cause, in adults and in children; it occurs more frequently in women (especially near the menopause) than in men. Myxedema is provoked by total, and sometimes by subtotal, thyroidectomy (postoperative myxedema); it is also a complication of certain infectious diseases (postinfectious myxedema). Prolonged and excessive treatment with iodine, radioactive iodine, or thiouracil and other antithyroid drugs may also produce myxedema. Secondary myxedema is observed in certain cases of severe anterior hypophyseal insufficiency, and in some patients suffering from Addison's disease (see Chap. 54).

Thyroid hyperfunction. There are three forms of hyperthyroidism in man: (*a*) exophthalmic goiter; (*b*) toxic adenoma of the thyroid; (*c*) hyperthyroidism without goiter. Exophthalmic goiter[1] (Fig. 247) has the following signs and symptoms: (*a*) a soft vascular goiter; (*b*) exophthalmos, *i.e.*, protruding eyes with widely open lids (in two-thirds of the cases); (*c*) tachycardia and sometimes palpitations; (*d*) tremor, which is fine and rapid; (*e*) a BMR of +20 or more; (*f*) muscular signs, such as incoordination of the eye and eyelid muscles (Graefe's sign), incomplete convergence of the eye (Möbius's sign), muscular atrophy and weakness, fatigue; (*g*) a tendency to lose weight; (*h*) increased respiratory frequency, diminished vital capacity, and occasionally dyspnea; (*i*) rapid circulation with a large minute volume and a short circulation time. If radioactive iodine is given, it is rapidly taken up by the thyroid, where it is accumulated in higher concentrations than in the normal thyroid. Urinary elimination in the first 48 to 72 hr. is subnormal. The protein-bound iodine in plasma is increased (more than 8 μg per 100 cc.).

Toxic adenoma is seen in subjects who have had a goiter for some time without functional disturbances, hyperthyroidism appearing later. Ocular signs are usually absent.

All forms of thyroid disease, but especially hyperthyroid conditions, are more frequent in women than in men.

[1] Known as "Parry's disease" or "Graves' disease" in England, "Basedow's disease" in Germany, and "Flaiani's disease" in Italy.

BIBLIOGRAPHY

ABELIN, I., Thyroid, in Bethe's "Handbuch der normalen und pathologischen Physiologie," **16,** 94, Springer, Berlin, 1930.

ALBERT, A., The Biochemistry of the Thyrotropic Hormone, *Ann. New York Acad. Sc.*, **50,** 466, 1949.

ASTWOOD, E. B., Chemotherapy of Hyperthyroidism, *Harvey Lect.*, **40,** 195, 1944.

ASTWOOD, E. B., Mechanism of Action of Antithyroid Compounds, *Ann. New York Acad. Sc.*, **50,** 419, 1949.

BARKER, S. B., Mechanism of Action of the Thyroid Hormone, *Physiol. Rev.*, **31,** 205, 1951.

CURTIS, G. M., and M. B. FERTMAN, Iodine in Nutrition, *J. A. M. A.*, **139,** 29, 1949.

DRILL, V. A., Interrelations between Thyroid Function and Vitamin Metabolism, *Physiol. Rev.*, **43,** 1943.

ELMER, A., "Iodine Metabolism and Thyroid Functions," Oxford, New York, 1938.

Endocrinology.

FERTMAN, M. B., and G. M. CURTIS, Foods and Genesis of Goiter, *J. Clin. Endocrinol.*, **11,** 1361, 1951.

HARINGTON, C. R., Twenty-five Years of Research on the Biochemistry of the Thyroid Gland, *Endocrinology*, **49,** 401, 1951.

HOUSSAY, B. A., "Acción de la Tiroides sobre el Metabolismo de los Hidratos de Carbono," El Ateneo, Buenos Aires, 1945.

Journal of Clinical Endocrinology.

MCGAVACK, T. H., "The Thyroid," Mosby, St. Louis, Mo., 1951.

MEANS, J. H., "The Thyroid and Its Diseases," Lippincott, Philadelphia, 1948.

NEW YORK ACADEMY OF SCIENCES, Thyroid Function as Disclosed by Newer Methods of Study, *Ann. New York Acad. Sc.*, **50,** 270, 1949.

PONSE, K., L'histophysiologie thyroidienne, *Ann. d'endocrinol.*, **12,** 266, 1951.

RAWSON, R. W., "Ciba Foundation Colloquia on Endocrinology," Vol. 4, 295, 1952 (thyrotropic hormone activity).

SALTER, W. T., The Chemistry and Physiology of the Thyroid Hormone, and the Control of Thyroid Activity, in Pincus, G., and K. V. Thimann, "The Hormones," Vol. 2, Academic Press, Inc., New York, 1949, pp. 181 and 301.

STANBURY, J. B., *et al.*, The Iodine Deficient Human Thyroid Gland, *J. Clin. Endocrinol.*, **12,** 191, 1952.

STARR, P., *et al.*, Clinical Experience with Blood Protein Bound Iodine Determination as a Routine Procedure, *J. Clin. Endocrinol.*, **10,** 1237, 1950.

Transactions of the American Association for the Study of Goiter.

CHAPTER 54

The Adrenal Glands

THE ADRENALS ARE endocrine glands which fulfill functions of vital importance, and total adrenalectomy is fatal in most vertebrates if adequate treatment is not given.

Historical note. The adrenals were discovered by Eustachius in 1563. Bartholinus (1611) attributed to them the secretion of the atrabile or black bile. Winslow gave them the name of "suprarenal glands," a term that is appropriate for the human adrenal but not for that of certain other species. The term "adrenal," which means "near the kidney," is now in common usage. Addison (1855) described the symptoms caused by destructive lesions of the adrenals, and Brown-Séquard performed experimental adrenalectomy for the first time in 1856. The blood-pressure-raising effect of adrenal extracts was reported by Oliver and Schäfer in 1894 and led to the discovery of adrenaline by Takamine and Aldrich in 1901. At first all the functions of the adrenals were supposed to be due to the secretion of adrenaline by the adrenal medulla, but by 1921 definite proof had been obtained that the cortex and not the medulla is essential for life. Since 1929 several active principles have been isolated from the adrenal cortex, their chemical structure has been established, and their physiologic significance is fairly well understood.

Physiologic anatomy. The adrenal glands of mammals consist of two parts, which differ by their origin, structure, and functions. The medulla arises in the ectoderm from cells that are also the origin of the sympathetic ganglia. These primitive cells, or sympathoblasts, become differentiated into the neurons of the sympathetic ganglia or into glandular cells which are called "chromaffin" (Henle, 1865) or "chromophil" (Stilling) cells because they stain a dark brown with chromates. The adrenal medulla is made up of these chromaffin cells. The medulla is completely surrounded by the cortex in mammals. The cortex has its origin in the mesoderm. The adrenal develops rapidly to a large size in the fetus. The cortex grows relatively faster than the medulla. Two zones can be clearly differentiated in the fetal cortex, an external and an internal layer. The latter retrogresses toward the end of fetal life and the first months after birth; by the end of the first year of life it has been completely reabsorbed. At birth the adrenals are still very large; their size is more than one-third that of the kidney. In the adult the adrenals are relatively much smaller; both weigh only 10 to 12 gm. in adult man.

In certain fishes, the *Elasmobranchii*, the two parts of the adrenal are not joined. In other vertebrate species (amphibians and reptiles), the tissues come together and are intermingled; only in mammals are they segregated into cortex and medulla. In mammals there are also small masses of chromophil cells outside the adrenals, known as paraganglia (Kohn). The most prominent of these masses is Zuckerkandl's aortic body, situated near the root of the inferior mesenteric artery, but others are found in the celiac plexus and in the neighborhood of the sympathetic ganglia. The so-called "accessory adrenals" usually are formed only by cortical tissue. Accessory adrenals are seldom found in certain species (cat, dog); in others they occur frequently (rabbit, white rat).[1] In man, accessory adrenals have been found in the kidney, liver, pancreas, epididymus, the broad ligament, and the neighborhood of the ovary.

The adrenals have a very active circulation; the minute volume varies from 5.3 to 6.4 cc. per gm. of gland.[2] The adrenals receive numerous sympathetic fibers from the solar and renal plexuses, which leave the spinal cord in the

[1] LASCANDO-GONZÁLEZ, J. M., *Rev. Soc. argent. de biol.*, **10**, 28, 1934.

[2] HOUSSAY, B. A., and E. A. MOLINELLI, *Rev. Soc argent. de biol.*, **2**, 117, 1926.

major and minor splanchnic nerves and in the rami of the lumbar roots. Most of these fibers end in the medulla. Some are vasomotor; others stimulate the secretion of adrenaline. The latter exert a continuous tonic effect on the chromophil cells; after they have been cut, adrenaline secretion is reduced to traces.[1] The adrenals also receive fibers from the vagus nerves, but their significance is unknown.

It has often been mistakenly affirmed that complete denervation of the adrenals leads to atrophy of the medulla. This is not so, because several weeks after complete denervation the chromophil cells not only have a normal aspect on histological examination, but respond with a copious discharge of adrenaline if they are stimulated directly (*e.g.*, by nicotine). Atrophy, when it occurs, is due to damage of the blood vessels in the course of the denervation.[2]

Three layers can be distinguished in the adrenal cortex; from without inward, these are the glomerulosa, fasciculata, and reticularis. Some workers maintain that the cells of the glomerular layer multiply and become differentiated into the cells of the other layers, but this opinion is not universally accepted. When there is hypertrophy or atrophy of the adrenals, changes in the inner layers are more marked than in the outer layers. The zona fasciculata is the widest of the three. It has polyhedral cells with many vacuoles filled with fat droplets—neutral fats, cholesterol esters, phospholipids, and specific sterols which are the active principles of the adrenal cortex. These cells reduce osmic acid because they contain a high percentage of oleic acid. The vacuoles give these cells a spongelike aspect; they are therefore known as spongiocytes. Cortical cells have a high ascorbic-acid content; this substance was obtained first from the adrenal cortex and only later from plant tissues (Szent-Györgyi, 1928).

INTERRELATION BETWEEN THE ADRENALS AND THE HYPOPHYSIS

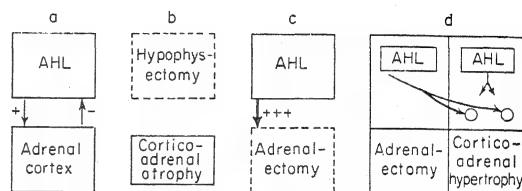
There is an important functional reciprocal relation between the pars distalis of the hypophysis and the adrenals.

The pars distalis secretes adrenocorticotrophin, which is one of the main factors in the

[1] Lewis, J. T., and R. O. Prieto, *Rev. Soc. argent. de biol.*, **14**, 555, 1938.

[2] Sgrosso, J. A., *Rev. Soc. argent. de biol.*, **11**, 139, 1935.

development and maintenance of corticoadrenal structure and function. Adrenocorticotrophin exerts its effect continuously; therefore hypophysectomy (or extirpation of the pars distalis) is rapidly followed by marked atrophy of the adrenal cortex. Corticoadrenal function is considerably diminished but not completely suppressed by hypophysectomy, since subsequent adrenalectomy causes a typical and fatal corticoadrenal insufficiency.

a: AHL; + −; Adrenal cortex
b: Hypophysectomy; Corticoadrenal atrophy
c: AHL; +++; Adrenalectomy
d: AHL; AHL; Adrenalectomy; Corticoadrenal hypertrophy

Adrenals and hypophysis: *a*, normal white rat; *b*, hypophysectomized rat; *c*, adrenalectomized rat; *d*, parabiosis between a normal and an adrenalectomized rat.

Corticoadrenal hormones regulate the secretion of hypophyseal adrenocorticotrophin. In the normal animal corticoadrenal hormones inhibit adrenocorticotrophin secretion; total adrenalectomy is followed by an increase in adrenocorticotrophin secretion, which is less marked and transitory if only one adrenal, or one adrenal and part of the other, are removed. This can be demonstrated by joining in parabiosis a normal and an adrenalectomized rat. The latter secretes an excess of adrenocorticotrophin, which passes into the intact partner and causes the adrenals to hypertrophy. By a similar mechanism, if the adrenals are removed from a pregnant rat, the adrenals of the fetus are seen to be hypertrophied at birth.[1] Adrenocorticotrophin has been assayed in the blood and found to be increased in adrenalectomized animals (Genzell, Snydor).

Compensatory hypertrophy, *i.e.*, hypertrophy of the fragment of adrenal tissue remaining after subtotal adrenalectomy, does not take place if the hypophysis has been removed. Total adrenalectomy in the rat is followed by hypertrophy of the accessory adrenals, which are found in this species. These accessory glands develop gradually into large masses, which at a certain stage of development pass through a phase of hyperactivity. The adrenals have a

[1] Houssay, B. A., and R. M. Pinto, *Rev. Soc. argent. de biol.*, **20**, 38 and 108, 1944; **21**, 316, 1945.

considerable capacity for regeneration in the rat, but it is not so marked in other species.

Changes in size and activity of the adrenals, in response to different factors, are almost always due to an increase or a decrease in adrenocorticotrophin secretion.

Agents which increase adrenocorticotrophin secretion act in the following ways: (*a*) by stimulating hypothalamic centers which control the hypophysis; (*b*) by direct stimulation of the hypophysis; (*c*) by releasing substances, *e.g.*, adrenaline, which stimulate the hypothalamus or the hypophysis (see Functions of the Hypophysis, Chap. 52).[1]

Adrenocorticotrophin was first obtained as a pure protein (Li *et al.*, 1943; Sayers *et al.*, 1943); then a small very active part was separated by adsorption, or chromatography, from a residue with little activity. Partial hydrolysis does not destroy the activity of adrenocorticotrophin, which is found in an ultrafilterable substance, apparently a polypeptide (Li). A fraction which diminishes ascorbic acid in the adrenal and another which causes adrenal hypertrophy have been separated from adrenocorticotrophin.

Assay of adrenocorticotrophin. Injection of adrenocorticotrophin provokes the following effects, which have been used to assay the potency of adrenocorticotrophin preparations: (*a*) increase in weight in the infantile or adult adrenal; (*b*) prevention of corticoadrenal atrophy or recovery of the adrenal size and aspect after hypophysectomy in rodents; (*c*) early reappearance of substances staining with sudan in the adrenal cortex of hypophysectomized rats; (*d*) rapid decrease of cholesterol and ascorbic acid in the adrenals. It would also be possible to assay corticoadrenotrophin by its effect on the secretion of corticoadrenal hormones by an isolated surviving perfused adrenal.

The most rapid and sensitive method of assaying adrenocorticotrophin preparations is the following: One of the adrenals is removed from hypophysectomized rats and the preparation to be assayed is injected; 1 hr. later, the second adrenal is removed. Ascorbic-acid concentration is estimated in both adrenals, and it will be found diminished in the second adrenal if the preparation contained adrenocorticotrophin.[2]

[1] HUME, D., G. HARRIS, J. DE GROOT, D. JACOBSOHN, and C. FORTIER, "Ciba Foundation Colloquia on Endocrinology," Vol. 4, J. & A. Churchill, London, 1952.

[2] SAYERS, G., and M. A. SAYERS, *Recent Progress in Hormone Res.*, **2**, 81, 1948; *Endocrinology*, **42**, 379, 1948.

Anatomic or functional changes in the organism should be attributed to corticoadrenal secretion provoked by adrenocorticotrophin if the following evidence is available: (*a*) the effect is observed in intact animals, but not in hypophysectomized or adrenalectomized animals; (*b*) the effect is obtained by injection of adrenocorticotrophin in intact or hypophysectomized, but not in adrenalectomized, animals. The following supplementary evidence may also be obtained: (*a*) there is an increase in the urinary elimination of substances with the activity of corticoadrenal hormones, or known to be derived from them; (*b*) one or more of the corticoadrenal hormones has the same effect in normal and adrenalectomized animals as the injection of adrenocorticotrophin.

The effects of adrenocorticotrophin are estimated in man indirectly by the effects produced by the excess of corticoadrenal hormones secreted, such as (*a*) rapid decrease of eosinophil cells in the blood and increase in uric acid excretion;[1] (*b*) decrease in the lymphocyte count in the blood; (*c*) increased excretion of substances in the urine which have corticoadrenal activity, such as a protective effect against the lethal action of cold on adrenalectomized animals[2] or the storage of glycogen in the liver of adrenalectomized mice; (*d*) estimation of the urinary elimination of steroids (17-ketosteroids, corticoids, or oxycorticoids) which can be derived from corticoadrenal hormones.[3]

INTERRELATION BETWEEN THE ADRENALS AND OTHER ENDOCRINE GLANDS

Thyroid gland. The thyroid gland stimulates the adrenals indirectly through the hypophysis. The administration of thyroid preparations, thyroxine, or thyrotrophin (if the thyroid is intact and capable of responding) provokes an increase in size and weight of the adrenals due to hypertrophy of the cortex. This effect is not obtained in hypophysectomized animals, except in the pigeon, in which the thyroid has a direct effect on the adrenals. Thyroidectomy is followed by a decrease in the size of the adrenals in the rat, and in patients suffering from hypo-

[1] THORN, G. W., *et al.*, *J. A. M. A.*, **137**, 1005, 1948.

[2] SELYE, H., and V. SCHENKER, *Proc. Soc. Exper. Biol. & Med.*, **39**, 518, 1938.

[3] MASON, H. L., *et al.*, *J. Clin. Endocrinol.*, **8**, 1, 1948.

thyroidism there is a depression of corticoadrenal function and the response to adrenocorticotrophin is diminished. Large doses of thyroxin, which inhibit hypophyseal function, cause hypofunction of the adrenal. Prolonged treatment with corticoadrenal steroids, *e.g.*, cortisone, transitorily depresses thyroid function.

Many of the effects of thyroxin, such as the rise in the BMR and nitrogen catabolism, are considerably diminished after adrenalectomy, but a normal response is obtained if corticoadrenal hormones are previously injected.[1]

Sexual glands. The adrenals are of larger size in the female than in the male owing to greater development of the cortex. The difference is particularly marked in the deeper layers of the zona fasciculata and the zona reticularis, especially in female mice, in which species there is a differentiated internal layer known as the X zone.

This sexual difference is due to the stimulating effect of female hormones and inhibitory effect of male hormones on the adrenal cortex. Thus castration causes a slight increase in the weight of the adrenals in the male and a slight decrease in the female.[2] The effect of female sexual hormones is carried out through the hypophysis and is suppressed by hypophysectomy. Prolonged treatment with large doses of estrogens strongly inhibits the hypophysis, thus causing atrophy of the adrenals (Zondek).

Certain strains of mice and rats develop corticoadrenal tumors after castration, which secrete male or female hormones having a strong stimulating effect on the sexual organs.

PHYSIOLOGIC AND PATHOLOGIC CHANGES IN THE ADRENALS

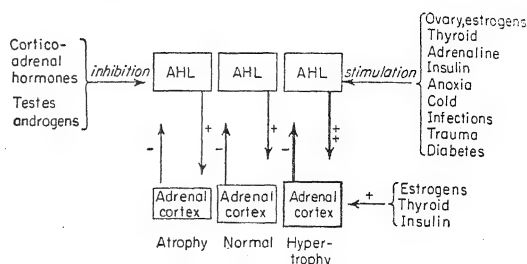
The adrenal gland responds readily to many physiologic and pathologic factors by changes in size, structure, and activity.[3] These changes are brought about by two mechanisms: (*a*) variations in the amount of adrenocorticotrophin secreted by the pars distalis of the hypophysis; (*b*) a direct effect on the adrenal cortex.

[1] Hoffmann, F., E. J. de Hoffmann, and J. Talesnik, *Rev. méd. de Chile*, **75**, 581, 1947; *J. Physiol.*, **107**, 251, 1948.

[2] Woolley, G. W., E. Fekete, and C. C. Little, *Cancer Research*, **3**, 120, 1943; **5**, 193, 203, 211, 321, and 506, 1945.

[3] Tepperman, J., F. L. Engel, and C. N. H. Long, *Endocrinology*, **32**, 403, 1943.

Cortico-adrenal hormones, Testes androgens — inhibition → AHL | AHL | AHL ← stimulation — Ovary, estrogens, Thyroid, Adrenaline, Insulin, Anoxia, Cold, Infections, Trauma, Diabetes

Adrenal cortex (Atrophy) | Adrenal cortex (Normal) | Adrenal cortex (Hypertrophy) ← + — Estrogens, Thyroid, Insulin

Adrenal Atrophy

Adrenal atrophy is due mainly to the decrease or suppression of hypophyseal secretion of adrenocorticotrophin.

Hypophysectomy. Removal of the hypophysis suppresses the source of adrenocorticotrophin; therefore it is followed by marked atrophy of the adrenal and a certain degree of adrenal insufficiency (Fig. 242). Atrophy is especially notable in the internal layers of the cortex. Signs of adrenal insufficiency have been reported in patients with hypophyseal insufficiency. Hypophysectomy suppresses adrenal response to the administration of thyroid, estrogens, and other agents which cause adrenal hypertrophy in animals with the hypophysis intact. Adrenal grafts do not "take" in hypophysectomized rats as readily as in normal animals

Adrenal hormones. Repeated injections of large doses of adrenal hormones inhibit the secretion of adrenocorticotrophin with the following results: (*a*) adrenal weight decreases because of atrophy of the cortex (see diagram) and there is adrenal hypofunction; (*b*) after unilateral adrenalectomy the remaining adrenal does not hypertrophy; (*c*) adrenal grafts do not "take." The principal effect of adrenal hormones is carried out through the hypophysis, but there is also a slight direct action on the adrenal cortex, which is made evident by the injection of adrenal hormones in hypophysectomized rats with atrophic adrenals; in this case the adrenals undergo further atrophy. In cases of adrenal tumor with cortical hyperactivity, the other adrenal is frequently atrophic and its function is subnormal.

Male hormones. Male hormones inhibit the adrenal cortex (see diagram). Castration in males is therefore followed by an increase in the size of the adrenals. Testosterone injections cause a decrease in the size of the adrenals and inhibit the secretion of corticoadrenal hormones. These effects are followed in man by a decrease in the

urinary excretion of glucocorticoids. Testosterone treatment in the rat, however, at first delays adrenal atrophy following hypophysectomy.

Adrenal Hypertrophy

Adrenal hypertrophy is caused by an increase in the secretion of adrenocorticotrophin by the pars distalis of the hypophysis. Only the cortex is involved in this process, which does not take place if the hypophysis has been removed. There is usually dilatation of the sinusoids, especially in the inner layers of the cortex. Sometimes the dilatation is so marked (*e.g.*, in guinea pigs that have been given a fatal dose of diphtheria toxin) that it has the appearance of a hemorrhage and is mistakenly called "adrenal hemorrhage." The increase in size is accompanied by functional hyperactivity of the cortex at least during certain stages of the process.

Excess adrenocorticotrophin. The administration of adrenocorticotrophin causes hypertrophy of the adrenal cortex. Total anterior hypophyseal extract also produces corticoadrenal hypertrophy, acting directly on the adrenal cortex by means of the adrenocorticotrophin it contains, and indirectly because thyrotrophin and gonadotrophins stimulate the secretions of the thyroid and sexual glands respectively, and the hormones of these glands stimulate adrenocorticotrophin secretion. The direct effect of adrenocorticotrophin can be demonstrated by injecting it into hypophysectomized rats, which respond with hypertrophy of the adrenal cortex.

Compensatory hypertrophy. Extirpation of one adrenal and part of the other is followed by hypertrophy of the fragment of adrenal tissue remaining in the animal. Bilateral adrenalectomy causes hypertrophy of the accessory adrenals. This compensatory response is due to an increase in the secretion of adrenocorticotrophin; it does not take place in hypophysectomized animals.

Female hormones. Injection of estrogens, or gonadotrophin (if the ovary is intact and functioning) is followed by corticoadrenal hypertrophy if the anterior hypophysis is capable of response, but not after hypophysectomy.

Thyroid hormones. Thyroid administration, or the injection of thyroxine or of thyrotrophin (if the thryoid is intact), causes corticoadrenal hypertrophy. This is not observed in hypophysectomized aniamls, except in the pigeon, in which the thyroid hormone acts directly on the adrenal cortex, even after thyroidectomy or hypophysectomy.

Metabolic factors. Excess protein in the diet, or an increase in protein catabolism, is a favorable factor for adrenal hypertrophy. The adrenals are found enlarged in animals with severe, but not in those with mild, diabetes. Malnutrition and inanition cause adrenal hypertrophy or, in severe cases, atrophy of the adrenals, which is nevertheless not so intense as the atrophy of other organs.

Accidental hypertrophy. Many physical, chemical, and biological (microorganisms) agents provoke adrenal hypertrophy (see The general adaptation syndrome, p. 622).

THE FUNCTIONS OF THE ADRENAL MEDULLA

The only well-demonstrated function of the adrenal medulla is the secretion of adrenaline and noradrenaline, which are synthesized from phenylalanine. Adrenaline is secreted continuously by a "tonic" action of cholinergic sympathetic fibers. Section of these fibers does not completely suppress this secretion, but reduces it to traces.

The adrenal medulla is not essential for life. It can be removed completely without causing disturbances in any of the main functions of the organism. The blood pressure, BMR, blood-sugar level, sympathetic responses, etc., do not vary significantly if the adrenal cortex remains intact.[1] Adrenaline secreted by the adrenal is not indispensable for the performance of these functions, but sympathin is still released at sympathetic nerve endings after the adrenal medulla has been extirpated. Sympathin is a mixture of varying amounts (0 to 100 per cent) of noradrenaline and adrenaline (Von Euler, Bacq). The fetal adrenal produces mainly noradrenaline; the adult adrenal produces more adrenaline.

The adrenal medulla forms part of the sympathicoadrenal system, and its secretion increases and diminishes together with the activity of the sympathetic nervous system. Adrenaline and noradrenaline secreted by the adrenal have the same effects as stimulation of the sympathetic fibers which release sympathin (adrenergic fibers), but they have a more widespread

[1] During a few days after the removal of the adrenal medulla there is slight corticoadrenal insufficiency, due to the surgical trauma caused by the extirpation of the medulla. This is rapidly compensated.

effect than the latter, which acts locally. Adrenaline secretion will be considered in detail after the study of the sympathetic nervous system (Chap. 85). Adrenaline also stimulates the secretion of corticoadrenal hormones.

THE FUNCTIONS OF THE ADRENAL CORTEX

Hormones secreted by the adrenal cortex regulate many important functions, among which the following are the best known:

1. *Salt and water metabolism.*
   a. Regulation of distribution of water, sodium, potassium, and chloride in the body. This is of vital importance to the organism.
   b. Regulation of renal excretion of water, sodium, potassium, and chloride.
   c. Regulation of blood volume and maintenance of arterial blood pressure.

2. *Carbohydrate, fat, and protein metabolism.*
   a. Regulation of the blood-sugar level, of glycogen concentration in the tissues, and of production and consumption of glucose.
   b. Increase in protein catabolism and in glyconeogenesis from protein.
   c. Action on fat metabolism.

3. *Neuromuscular function.*
4. *Action on sexual functions*, the development of secondary sexual characters, hair growth, and skin.
5. *Resistance to many physical and chemical agents and to infections.*
6. *Interrelation with other endocrine glands and tissues* (thymus, lymphoid tissue, etc.).

Corticoadrenal Hormones

When it was found that death following adrenalectomy was due to cortical insufficiency, further experiments led to the discovery of cortical extracts which could maintain life in adrenalectomized animals.[1] Up to now 28 steroids have been separated from these extracts, though it is not known whether all of them are produced by the adrenal or whether some arise in the process of extraction.[2] The majority of these steroids are derivatives of allopregnane; they belong to the group of substances known as the perhydrocyclopentenephenanthrene ring system[1] and therefore are closely related to the sexual hormones. Six of the adrenal steroids have physiologic activity. The first to be identified was called corticosterone, and others were given names derived from this term or called by letters (Kendall), *e.g.*, compound E (cortisone).

Only 10 per cent of the active substance has been obtained in pure crystalline form and identified chemically. The remaining 90 per cent is the more active part, but its chemical structure has not yet been established.

Some of the steroids have a *masculinizing* effect, *e.g.*, adrenosterone, androstenedione, 11-hydroxyisoandrosterone, and 17-hydroxyprogesterone. Others have *feminizing* activity, *e.g.*, estrone (secreted by the ovary), or *progestational* activity, *e.g.*, progesterone (the hormone of the corpus luteum) and desoxycorticosterone. The adrenal gland also contains ascorbic acid and large quantities of sulfur.

The most important of the corticosteroids is 17-hydroxy-corticosterone, because it is the predominant one found in adrenal blood and has the greater number of physiologic activities. Desoxycorticosterone has been found only in a few cases in adrenal extracts, although it appears in the blood perfused through an isolated adrenal; there is therefore some doubt that it is a hormone produced by the normal gland. It is widely used because it maintains adrenalectomized animals and is efficacious in Addison's disease (chronic adrenal insufficiency). Aldosterone,[2] at first called electrocortin, has been isolated recently; it has a potent effect on sodium retention, between 25 and 50 times that of desoxycorticosterone. Cortisone (compound E, 11-dehydro-17-hydroxy-corticosterone) is frequently used in therapeutics, since Hench and his associates observed that it produced spectacular improvement in rheumatoid arthritis. Both these substances are prepared by synthesis from plant steroids or sapogenins (soia, Dioscoridae, or Strophanthus), from ox bile, or by the action of microbes.

[1] Between 1929 and 1931 Stewart and Rogoff, Hartman, and especially Swingle and Pfiffner obtained these extracts by different methods.

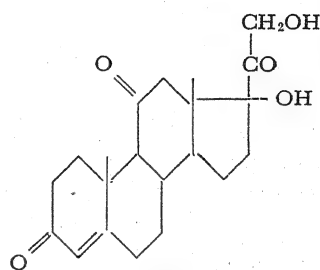
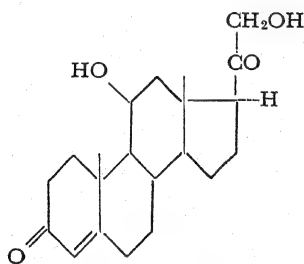
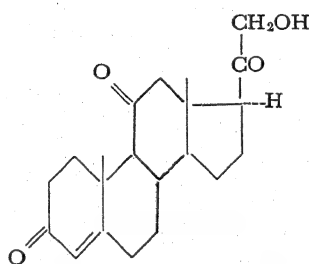
[2] Mainly by Reichstein, Kendall, Wintersteiner, and Pfiffner.

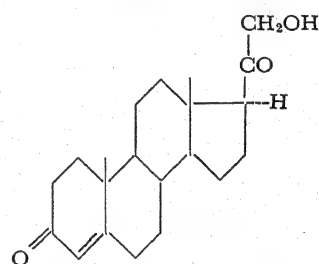
[1] For steroid nomenclature see Selye, H., "Textbook or Endocrinology," Acta Endocrinologica, University of Montreal, 1947. Mason, H. L., *J. Clin. Endocrinol.*, **8**, 190, 1948.

[2] Simpson, S. A., J. F. Tait, A. Wettstein, R. Neher, J. v. Euw, and T. Reichstein, *Experientia*, **10**, 132, 1954.

Corticosteroids in the body seem to derive from cholesterol. The perfused isolated adrenal produces, or transforms, several corticosteroids; *e.g.*, it substitutes an OH for one H on C^{11}, and desoxycorticosterone is converted to corticosterone; cortisone is converted into hydrocortisone by substitution of OH on C^{17}; progesterone is converted into several corticosteroids (Hechter *et al.*).

Allopregnane

17-hydroxycorticosterone (Hydrocortisone)

17-hydroxy-11-dehydrocorticosterone (Cortisone)

Corticosterone

Dehydrocorticosterone

Desoxycorticosterone

The following molecular structure is essential for great activity: a ketone on C^3, a double bond between C^4 and C^5, and a lateral chain of $CO \cdot CH_2OH$ on C^{17}. Corticosteroids with an oxygen atom or a βOH on C^{11} are active in carbohydrate metabolism (Thorn), retard the onset of fatigue (Ingle), increase protein catabolism (Long), and provoke reabsorption of lymphoid tissue cells (Dougherty, White). The elimination of oxygen from C^{11} suppresses almost completely the activity on carbohydrate metabolism but increases the effect on sodium and chloride retention (*e.g.*, desoxycorticosterone). Addition of OH on C^{17} increases the effect on carbohydrate metabolism but decreases the retention of Na and Cl and, in some cases, provokes loss of these elements (Thorn). Only cortisone (compound E) and hydrocortisone (compound F) are efficacious in rheumatoid arthritis.

Approximate quantitative effects of several corticosteroids and the amorphous fraction are given in Table 77.

According to their therapeutic effect, cortico-adrenal hormones have been classified in three groups:

1. Corticoids with activity in mineral metabolism, *e.g.*, aldosterone, desoxycorticosterone, and Reichstein's compound S, which provoke retention of Na, Cl, and water and an increase in the excretion of potassium and nitrogen.
2. Glycocorticoids (with oxygen on C^{11}), which increase the fasting blood-sugar and liver-glycogen concentration, increase diabetic hyperglycemia and glycosuria, stimulate glyconeogenesis from protein, increase protein catabolism, and decrease glucose consumption.
3. Androgenic corticoids, which increase muscular development and hair growth, provoke virilization in women (hypertrophy of the clitoris, hirsutism, etc.), and increase urinary excretion of 17-ketosteroids.

The difference between corticoids active on mineral metabolism and glycocorticoids is only relative; thus 17-hydroxycorticosterone has both activities (Table 77).

Secretion of Corticoadrenal Hormones

In adrenal blood (dog, ox, man) mainly 17-hydroxycorticosterone has been found, but

there is also corticosterone and other corticosteroids that are so far unidentified.

Vogt[1] has measured the amount of corticoadrenal hormones in blood of the adrenal vein by testing its power of increasing the resistance to cold of adrenalectomized rats. She found extraordinarily large quantities, equivalent to 17 kg. of adrenaline per day for a 10-kg. dog (the equivalent of 0.6 gm. of gland per kg. per min. from each gland). Other indirect methods have given similar results. Injection of adrenaline increases corticoadrenal hormone secretion. The hormone disappears rapidly from the blood, and it cannot be estimated in arterial blood.

Excretion. Urinary extracts with "corticoid" effects have been prepared. These extracts have an activity similar to that of the adrenal hormones, such as (*a*) maintenance of survival in adrenalectomized animals; (*b*) increased resistance to cold; (*c*) increased glycogen. In cases of adrenal insufficiency this activity disappears. Part of the 17-ketosteroids in the urine comes from the adrenal, and patients suffering from Addison's disease excrete smaller amounts of these substances than normal subjects.

The daily elimination of corticoids (and oxycorticoids) increases in the following circumstances[2] (*a*) surgical operations, burns, cold, hemorrhage, flying at great heights, etc.; (*b*) Cushing's disease. It decreases in (*a*) Addison's disease; (*b*) hypophyseal insufficiency; (*c*) treatment with testosterone. In cases of hirsutism usually a normal amount of corticoids is eliminated in the urine. Urinary elimination of 17-ketosteroids increases (*a*) in the alarm reaction (transitorily); (*b*) in many cases of corticoadrenal tumors or hyperplasia; (*c*) usually in cases of hirsutism; (*d*) in some cases of Cushing's disease, but not in all; (*e*) following injection of adrenocorticotrophin. More than 60 steroids so far have been extracted from urine, and many of them have been chemically identified.[1]

Table 77. Approximate Quantitative Effects of Known Active Principles of the Adrenals*

Active principle	*Maintenance of life (rat)*	*Renal function (dog)*	*Electrolyte metabolism (dog)*	*Muscular contraction (rat)*	*Hepatic glycogen (rat)*	*Diabetogenic effect (rat)*
Corticosterone	1–2	2	1–2	3	3	3–4
11-Dehydrocorticosterone (compound A)	1–2	2	1–2	2	2	3–4
11-Dehydro-17-hydroxycorticosterone (compound E, cortisone)	2	1	?	4	4	5
17-Hydroxycorticosterone (compound F)	2	1	?	5	5	5
11-Desoxycorticosterone	4	4	5	1	1	1
Amorphous fraction	5+	5	?	1–2	1	?

Source: INGLE, D. J., "A Symposium on Steroid Hormones," University of Wisconsin Press, Madison, 1950.
* Activity is graduated from 1 (slight) to 5 (marked).

ADRENAL INSUFFICIENCY

Total extirpation of the adrenals is followed by a short period in which there are no disturbances and the animal appears normal. Later signs and symptoms gradually begin and increase progressively until the animal dies (Fig. 248). In the dog and cat an early sign is loss of appetite, followed by weakness, and the animal easily tires. Later digestive disturbances become evident, *e.g.*, vomiting, and diarrhea streaked with blood. There is difficulty in standing and walking, the blood pressure begins to fall, and death occurs with symptoms of shock or, more seldom, in convulsions.

Vital importance of the adrenals. Total adrenalectomy is followed by death within a short time in most species. This is due to cortical insufficiency, not to the loss of the medulla. Death is usually due to disturbances in water and electrolyte metabolism, and can be pre-

[1] VOGT, M., *J. Endocrinol.*, **5**, 57, 1945.

[2] VENNING, E. H., and J. S. L. BROWNE, *J. Clin. Investigation*, **25**, 935, 1946; FORBES, A. P., *et al.*, *J. Clin. Endocrinol.* **7**, 264, 1947; SHIPLEY, R. A., *et al.*, *J. Clin, Investigation*, **25**, 673, 1946; PINCUS, G., *Recent Progress in Hormone Res.*, **1**, 123, 1947; *Ann. New York Acad. Sc.*, **50**, 635, 1949.

[1] LIEBERMAN, S., K. DOBRINER, B. R. HILL, L. F. FIESER, and C. P. RHOADS, *J. Biol. Chem.* **172**, 263, 1948; MILLER, A. M., and R. I. DORFMAN, *Endocrinology*, **46**, 514, 1950.

vented by the administration of sodium chloride, corticoadrenal hormones, or adrenal grafts. Less frequently death is due to hypoglycemia or circulatory collapse.

Unilateral adrenalectomy is well tolerated; it does not provoke any disturbances. In the dog it is possible to remove from five-sixths to seven-eighths of the adrenal tissue without fatal results, but total bilateral adrenalectomy is always fatal in fishes, amphibians, reptiles, birds, and mammals. Most animals survive for only a few days: cats, 6 to 10 days; dogs, 2 to 16 days. A survival of less than 24 hr. is usually due to surgical trauma or the anesthetic; according to Swingle, local anesthesia of the sympathetic ganglia near the adrenals prevents these early deaths. Birds die soon after adrenalectomy, and toads (*Bufo arenarum*) in 2 to 4 days. The ferret survives several weeks (Hartman). In rabbits accessory adrenals are frequently found, and approximately 20 per cent of the animals survive indefinitely. Rats differ in their resistance to adrenalectomy according to the strain from which they come. Some strains have a high mortality; in others, a high percentage survive.[1] The latter show signs at first of corticoadrenal insufficiency, which gradually retrogress as the accessory adrenals, which are found in these strains, hypertrophy (Lascano-González).

Death is due to extirpation of the cortex, not to that of the medulla. This is demonstrated by the following facts: (*a*) extirpation of the interrenal bodies in fishes, which are the equivalent of the adrenal cortex, is always fatal (Biedl); (*b*) extirpation of the adrenal medulla is not followed by symptoms of adrenal insufficiency unless the cortex is severely damaged,[2] but removal of the remaining cortex is always followed by a typical and fatal syndrome of adrenal insufficiency; (*c*) cases of atrophy of the adrenal cortex with symptoms of adrenal insufficiency, but without changes in the medulla, have been reported; (*d*) adrenaline does not relieve the symptoms of adrenal insufficiency or prevent death; (*e*) corticoadrenal grafts that "take" permit survival after double adrenalectomy; (*f*) active corticoadrenal extracts and some of the steroids extracted from the adrenal cortex relieve the symptoms of adrenal insufficiency and keep adrenalectomized animals alive as long as they are treated with these extracts.

[1] LEWIS, J. T., *Am. J. Physiol.*, **64**, 503, 1923.

[2] Wheelon and Vincent, 1917; Houssay and Lewis, 1921; Crowe and Wislocki, 1924; and others. During a few days following the operation there may be a slight degree of adrenal insufficiency (Foglia and Gerschman).

The vital importance of the adrenal glands is due mainly to the effects of their secretion on the regulation of water and salt metabolism.

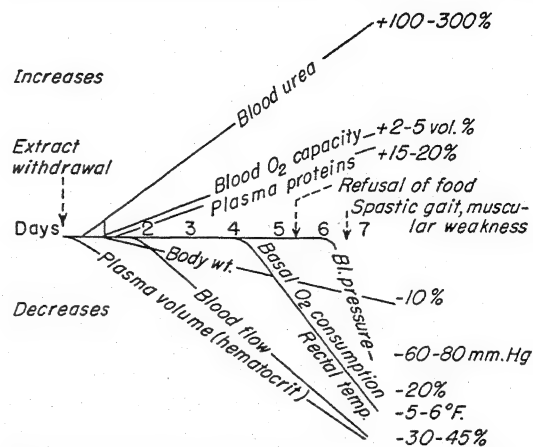



FIG. 248. Adrenal insufficiency. Characteristic changes in metabolism, circulation, and blood concentration in adrenalectomized animals following withdrawal of extract. (*Loeb, R. F., J. A. M. A., vol. 104, p. 2177, 1935.*)

The administration of sodium chloride alone will keep adrenalectomized animals alive. As stated, death is occasionally due to hypoglycemia or circulatory collapse. The adrenals are not indispensable to life if the disturbances in salt and water and carbohydrate metabolism are controlled.

Subtotal or even total adrenalectomy has been performed in man[1] in cases of cancer of the adrenal or the prostate, malignant hypertension, etc. The patients were kept alive by treatment with salt (5 to 10 gm. daily), desoxycorticosterone (0.1 to 1 mg. daily), and cortisone (25 to 50 mg. daily).

Disturbances in electrolyte metabolism. Adrenal hormones control reabsorption of sodium, potassium, and water by the renal tubes. In adrenal insufficiency sodium is lost in the urine because its reabsorption from the glomerular filtrate is diminished. Together with Na, Cl (but in smaller amounts) and water are also lost. Blood plasma, thus, has less sodium, chloride, bicarbonate, and water; the blood becomes

[1] THORN, G., *et al.*, *J. Clin. Endocrinol.*, **11**, 774, 1951; BERGENSTAL, D. M., and C. B. HUGGINS, *J. Clin. Investigation*, **31**, 616, 1952; PEARSON, O. H., *et al.*, *J. Clin. Investigation*, **31**, 653, 1952.

concentrated; and there is a relative increase in erythrocytes, hemoglobin, and plasma protein. Sodium, chloride, and water also diminish in extracellular fluids; there is dehydration, the blood volume diminishes, and a tendency to shock develops.

Potassium reabsorption increases simultaneously with the decrease in sodium reabsorption. Less K is excreted in the urine and its concentration in plasma increases. Intracellular fluid loses sodium, its water content increases, and K concentration rises or remains normal.

The concentration of potassium in plasma may reach toxic levels, and the severity of the effects of adrenal insufficiency becomes greater (Zwemer). The administration of K clearly increases the intensity of these effects, and may cause them to appear in patients with compensated chronic adrenal insufficiency.

Baumann (1927) observed the decrease in blood sodium, and Loeb (1932) demonstrated its importance. Adrenalectomized animals can be kept alive by giving them sodium (usually NaCl). If K in the diet is reduced, sodium is more efficacious, *e. g.*, K may be diminished from 4 gm. per day to 2 gm.

Corticoadrenal extract and several corticosteroids diminish the loss of Na and increase the excretion of K in adrenalectomized animals, because they increase reabsorption of Na and diminish that of K. Aldosterone and desoxycorticosterone are the most active; corticosterone, 11-dehydrocorticosterone, and 17-hydroxycorticosterone follow in activity.

Aldosterone is the most potent sodium-retaining steroid so far discovered; it has been found in adrenal venous blood. Other steroids have only a brief effect, or may increase sodium excretion, *e.g.*, cortisone. Desoxycorticosterone maintains adrenalectomized animals alive and is efficacious in patients suffering from adrenal insufficiency.

Disturbances in water metabolism. The adrenals exert an indirect effect on water metabolism by their activity on electrolyte reabsorption by the renal tubes, but they also have a direct effect on renal excretion of water. Following adrenalectomy, in a few species, there is increased elimination of water simultaneously with the loss of Na and Cl. This increase persists in animals treated with sodium chloride. When, however, severe insufficiency develops, the amount of urine eliminated diminishes, especially when there is dehydration. A typical disturbance is the slow rate of elimination of ingested water in adrenalectomized animals and in patients with Addison's disease. This disturbance, together with the decrease in NaCl in the body, increases sensitiveness of adrenalectomized animals toward water intoxication. The deficiency in the elimination of water is mainly of renal origin, as is shown by the following facts: (*a*) glomerular filtration decreases owing to diminished renal blood flow, caused by lower blood pressure and speed of circulation; (*b*) tubular reabsorption of water increases, this factor seeming to be the principal one (Gaunt); (*c*) sensitiveness to the posthypophyseal antidiuretic hormone increases, and there is a higher concentration of this hormone in the blood plasma (Gaunt). There is some evidence, but no definite proof, of an extrarenal mechanism.

Corticoadrenal hormones control these disturbances by their effects on electrolyte metabolism and on the kidney. Desoxycorticosterone, adrenocorticotrophin, and large doses of cortisone may produce edema by retention of NaCl which causes an increase in extracellular and a decrease in intracellular fluid. Thirst is usually increased (especially when desoxycorticosterone is administered) and the excretion of urine rises. In certain conditions prolonged treatment with desoxycorticosterone—or, in the dog, with cortisone—provokes intense polyuria similar to diabetes insipidus.

According to Gaunt, hormonal control of water excretion by the kidney is governed by (*a*) the neurohypophysis (antidiuretic hormone), and (*b*) the pars distalis of the hypophysis (adrenocorticotrophin, somatotrophin, and thyrotrophin). The neurohypophyseal hormone and the adrenal cortex apparently exert antagonistic effects, and the renal excretion of water depends on the maintenance of an equilibrium between them.

Sodium chloride maintains adrenalectomized animals alive, but does not compensate for all the effects of adrenal insufficiency, *e.g.*, disturbances in carbohydrate metabolism.

Adrenal insufficiency is not due exclusively to the loss of NaCl and water in the urine. Circulatory collapse and other symptoms of adrenal insufficiency are seen in adrenalectomized animals in which the kidneys have been removed. Cortical extracts prolong the survival of these animals. Moreover cases of

Addison's disease have been reported in which collapse and death occurred without any disturbance in plasma-electrolyte concentration.

Renal disturbances. Renal functions are gradually impaired as adrenal insufficiency increases in severity. At first there is a deficient reabsorption of Na, which is excreted in excess in the urine, and the body loses not only Na, but also Cl and water. Dehydration, hemoconcentration, and loss of body weight follow. Later K and water are reabsorbed in excess, and the urinary excretion of K and water diminishes. Urea and creatinine clearance diminishes, as also the clearance of other substances. The decrease in diuresis causes retention and an increase in blood plasma of several of its components: nonprotein nitrogen, urea, potassium, phosphates, and sulfates. Renal function is improved in adrenal insufficiency by several corticosteroids; the most active are the total extract, the amorphous fraction, and desoxycorticosterone.

Circulatory disturbances. The following disturbances in the circulation are seen in adrenal insufficiency: (*a*) a decrease in circulating blood volume, the size of the heart, and the speed of circulation; (*b*) a fall in blood pressure, especially marked and rapidly progressive during the phase of shock; (*c*) increased capillary permeability during shock. Adrenalectomized animals and patients suffering from adrenal insufficiency are very susceptible to all factors that provoke shock, such as trauma, hemorrhage, anesthesia, hypertonic glucose injections, or excessive heat or cold; exercise, and sometimes even the administration of a saline purge will provoke an adrenal crisis.

Adrenalectomy causes a fall in blood pressure and of hypertensinogen in the blood of animals suffering from hypertension provoked by renal ischemia. This fall in blood pressure is prevented by the injection of corticoadrenal hormones, but not by the administration of sodium chloride. Vascular response to several vasoconstrictor substances is diminished in adrenal insufficiency,[1] and when it is severe stimulation of sympathetic vasoconstrictor nerves is also followed by a subnormal response. Treatment with several corticosteroids improves vascular reactivity. These and other facts have led to the practice of total or subtotal adrenalectomy in the treatment of patients with malignant hypertension.

According to Swingle, Hartman, and others, changes in capillary and cell permeability are of special importance in adrenal insufficiency. Injections of saline solutions or blood transfusions do not raise the blood pressure of adrenalectomized animals, because the injected fluid passes out of the blood vessels into the tissues and edema is produced. Corticoadrenal extracts prevent this loss of fluid from the blood vessels (Swingle). Another demonstration of increased capillary permeability in adrenal insufficiency is obtained by joining in parabiosis a normal and an adrenalectomized rat. The latter soon becomes congested and edematized, owing to the higher blood pressure of the normal partner and increased capillary permeability in the adrenalectomized animal (Pinto and Houssay). Adrenal extract also inhibits or controls increased capillary permeability produced by leukotaxin (Menkin). Adrenal extracts and steroids have been used for the prevention and treatment of shock with good results in certain forms of experimental shock, but with less constant and satisfactory results in man.

[1] Cleghorn, R. A., *et al.*, *Am. J. Physiol.*, **161**, 21, 1950.

Digestive disturbances. In the course of acute adrenal insufficiency, the following signs of disturbance in digestive functions are observed: (*a*) anorexia; (*b*) vomiting; (*c*) diarrhea, sometimes streaked with blood. Intestinal absorption of sugars and fat is retarded in the adrenalectomized rat, but returns to normal if the animal is given sodium chloride.

Disturbances in carbohydrate metabolism. These have been fully discussed in Chap. 41, Carbohydrate Metabolism. In brief the principal disturbances observed in adrenal insufficiency are (*a*) tendency to hypoglycemia in fasting; (*b*) hypersensitiveness to hypoglycemic agents (insulin, phlorhizin, fasting, etc.); (*c*) marked decrease in the fasting level of hepatic and muscle glycogen; (*d*) disturbance in the selective intestinal reabsorption of sugar; (*e*) increased consumption of glucose by the adrenalectomized rat; (*f*) decreased severity of pancreatic and phlorhizin diabetes.

The administration of total cortical extract, adrenocorticotrophin, or the so-called "glycocorticoids" has the following effects in adrenalectomized animals: (*a*) a normal blood-sugar level is maintained; (*b*) glycogen, especially liver glycogen, increases; (*c*) insulin resistance increases; (*d*) sugar tolerance diminishes; (*e*) glucose consumption diminishes when large doses of corticosteroids are given; (*f*) diabetes attenuated by adrenalectomy or hypophysec-

tomy increases in severity; (*g*) large doses (far above those used in therapeutics) may provoke a transitory diabetes in the rat, and therapeutic doses increase the severity of an existing diabetes in man.

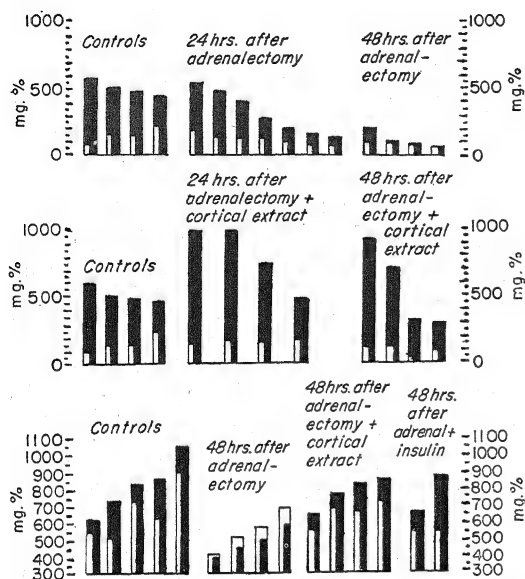



FIG. 249. Effect of adrenal cortex on muscle glycogen. Each column represents one experiment on a chloralosed dog. Two upper series: white, glycogen after fatigue; black, glycogen after 1 hr. recovery. Recovery is retarded in some animals 24 hr. after adrenalectomy and in all animals 48 hr. after adrenalectomy. Corticoadrenal extract accelerates recovery; in some cases, muscle-glycogen concentration is above that of normal untreated controls. Lower series: white, muscle glycogen before intravenous injection of 2 gm. glucose per kilogram; black, after the injection. Glycogen formation is deficient in adrenalectomized animals. Corticoadrenal extract and insulin restore the normal rate of muscle-glycogen formation in adrenalectomized animals. (*Dambrosi, E. G., L F. Leloir, and A. Novelli, Rev. Soc. argent. de biol., vol. 9, pp. 408 and 417, 1933.*)

Disturbances in protein and fat metabolism. Adrenalectomized animals usually do not grow well, and lose weight. In conditions of emergency protein catabolism does not increase in adrenalectomized as in normal animals, *e.g.*, in diabetes (Long and Lukens), following fractures and burns (Ingle), after the administration of thyroxin,[1] etc. Protein catabolism increases in all these circumstances if the animals are treated with small doses of corticosteroids; a basal level, not an increase in corticoadrenal secretion, is necessary.[1]

[1] HOFFMANN, F., *et al.*, *Bol. Soc. biol. Santiago, Chile*, **7**, 60, 1950.

In vitro experiments have shown that tissue slices taken from adrenalectomized animals have a subnormal deaminating activity when certain amino acids are used as substrates.

Adrenalectomy diminishes ketonuria in experimental diabetes, which may be a sign of a lower catabolism of fats. It is difficult to provoke fatty infiltration of the liver in adrenalectomized animals, possibly because of a decrease in the mobilization of storage fat toward the liver. Intestinal absorption of fats is retarded in adrenalectomized animals, and storage fat is considerably diminished (anorexia).

Other metabolic disturbances. The basal metabolic rate is well below normal (−25 per cent) in the acute stages of adrenal insufficiency, and oxygen consumption *in vitro* of surviving tissue slices taken from adrenalectomized animals is lower than that of normal tissues. Thyroxine has less effect on the BMR of adrenalectomized rats than in the controls.

Muscular disturbances. Addison, in his classical description of human adrenal insufficiency, reported the loss of strength (asthenia) and the rapidity of the onset of fatigue as a typical sign of the disease. Susceptibility to fatigue has been clearly demonstrated in adrenalectomized animals. This disturbance is controlled in part by the administration of glucose and completely by treatment with glycocorticoids which have oxygen on C^{11}, such as (in decreasing order of activity) 17-hydroxycorticosterone, cortisone, corticosterone, and 11-dehydrocorticosterone; 11-desoxycorticosterone has little activity.

In man adrenal insufficiency is accompanied by psychic depression.

Metabolism of pigments. An outstanding symptom of Addison's disease is a dark pigmentation of the skin, owing to an increase in the melanin of the epidermis. Nevertheless in acute or chronic adrenal insufficiency in experimental animals little or no disturbance in pigmentation has been observed.

The thymus and lymphoid tissue. The adrenal glands have an inhibitory effect on the thymus, the lymphatic glands, and lymphoid tissue. Adrenalectomy is followed by hypertrophy of the thymus and lymphatic glands.[2] Ad-

[1] INGLE, D. J., *Ann. New York Acad. Sc.*, **50**, 576, 1949.
[2] PINTO, R. M., *Am. J. Physiol.*, **144**, 652, 1945.

ministration of corticoadrenal hormones causes a decrease in size in these organs, which is permanent in some cases.

Adrenocorticotrophin injection is followed after a few hours (if the adrenals are intact) by marked involution of the thymus and lymphoid tissue, with widespread reabsorption of lymphoid cells.[1]

Decrease in the resistance of the organism. Adrenalectomized animals are very sensitive to factors that cause stress, and in situations of emergency they are easily exhausted, falling into collapse which is frequently fatal. This is especially evident when these animals are exposed to cold (Marval), or submitted to conditions that cause a fall in blood pressure (hemorrhage, histamine, trauma, etc.) or a decrease in blood sugar (insulin, phlorhizin). These animals are also sensitive to most drugs, poisons, and toxins, such as morphine, cobra venom, diphtheria toxin,[2] and potassium (Zwemer, Cicardo).

Injection of corticoadrenal extracts, or corticosteroids, especially glycocorticoids, increases the resistance of adrenalectomized animals, but it has little or no effect on the resistance of normal animals.

In the last century it was commonly believed that the effects of adrenal insufficiency were due to the accumulation of toxic substances, normally produced by the organism and neutralized by the adrenals,[3] but the toxic substances were never found. Later the metabolic disturbances of adrenal insufficiency and their control by the administration of corticoadrenal hormones or NaCl were discovered, and the toxic theory of adrenal insufficiency was definitely abandoned.

The role of the adrenals in sexual functions and in lactation will be discussed in the chapters on those subjects.

Treatment of adrenal insufficiency. Adrenalectomized animals can be kept alive indefinitely by the following methods: (*a*) injection of active corticoadrenal extracts or steroids; (*b*) a diet with a high sodium chloride but a low potassium content; (*c*) corticoadrenal grafts.

[1] RAPELA, C. E., *Rev. Soc. argent. de biol.*, **20**, 423, 1944; DOUGHERTY, T. F., and A. WHITE, *Endocrinology*, **35**, 1, 1944; **39**, 370, 1946; *Am. J. Anat.*, **77**, 81, 1945; *J. Lab. & Clin. Med.*, **32**, 584, 1947.

[2] LEWIS, J. T., *Am. J. Physiol.*, **64**, 506, 1923.

[3] Brown-Séquard, 1858; Abelous and Langlois, 1892; Langlois, 1897, and, more recently, Riml.

Certain species in which accessory adrenals are frequently found (*e.g.*, the rat and rabbit) survive indefinitely if they are adequately treated during the first few weeks after adrenalectomy, so that the accessory adrenals have time to develop sufficiently to replace the corticoadrenal tissue that has been removed.

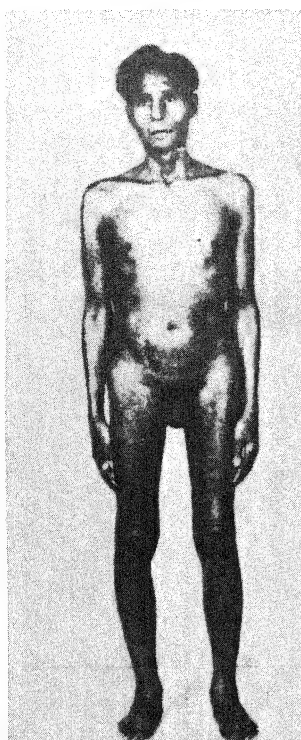
Immediately after adrenalectomy, adrenal extracts or hormones should be given together with NaCl. Later the dose of extract can be decreased, and finally it is possible to maintain the animals in good condition with sodium chloride alone. If the treatment with salt or extract is discontinued, the animals fall into acute adrenal insufficiency, unless accessory adrenals have developed. Extracts are more efficient than salt in the treatment of acute adrenal insufficiency. Adrenalectomized rats can be kept alive permanently, or their survival can be prolonged, by the administration of NaCl. If they can choose freely between pure water and 1 per cent NaCl solution they will prefer the latter (Richter). This is due to a lowered threshold for salt taste caused by the decrease in sodium in the blood and tissues.

There is apparently a certain antagonism between the effects of sodium and those of potassium in adrenalectomized animals. Sodium chloride is more efficient if there is little potassium in the diet; on the contrary an increase in K diminishes the effect of sodium and may cause acute adrenal insufficiency in an animal or a patient in which adrenal insufficiency has been previously compensated.

It should be noted that corticoadrenal grafts do not "take" easily in most species.

Corticoadrenal extracts improve the condition of adrenalectomized animals and restore them to a normal state. Renal reabsorption of Na increases, less sodium and chloride are excreted, and the urinary excretion of potassium increases. The concentration of sodium and chloride in plasma increases; that of potassium decreases. Plasma volume and arterial blood pressure return to normal levels. No single corticoadrenal steroid produces all the effects of the total extract. The activity of these extracts is due to the steroids they contain; therefore their effects and potency vary according to the method of preparation.[1]

Desoxycorticosterone (usually the acetate dis-

[1] Usually the potency of an extract is measured by its capacity to maintain alive totally adrenalectomized animals. It would be better to express their potency in terms of each one of the active steroids they contain.

solved in oil) is widely used in therapeutics. This substance controls the disturbances in salt metabolism and permits indefinite survival (Fig. 251). It causes an increase in the reabsorption of sodium in the renal tubes and therefore diminishes the loss of sodium, chloride, and water. It increases sodium and chloride and diminishes potassium in blood. It has little or no effect on carbohydrate metabolism, and adrenalectomized animals treated with desoxycorticosterone are liable to have hypoglycemic crises, which may prove fatal if they do not receive a sufficient amount of carbohydrate.

Large doses of desoxycorticosterone, especially if given together with salt treatment, may have unfavorable and even dangerous results. Plasma sodium and chloride increase above the normal, and in spite of polyuria there is a tendency to retain water. In the more advanced stages of intoxication by desoxycorticosterone, the following disturbances are observed: (*a*) marked subcutaneous edema, with accumulation of sodium chloride; (*b*) venous hypertension, dilatation of the heart, and sometimes fatal pulmonary edema; (*c*) arterial hypertension (not in all cases); (*d*) excessive decrease in plasma potassium, sometimes accompanied by cardiac disturbances and paralysis, in which case the potassium content of muscle is diminished and the sodium content increased; administration of potassium salts controls these disturbances. These toxic disturbances are prevented by keeping the dose of desoxycorticosterone below 5 mg. per day (except when treating an adrenal crisis), by keeping the ingestion of sodium chloride between 3 and 6 gm. per day without reducing the potassium intake, and by giving an adequate amount of carbohydrate to prevent hypoglycemia. If there is a sudden increase in weight (edema) the dose should be reduced immediately.[1]

Prolonged treatment with desoxycorticosterone produces in some animals signs of chronic intoxication, with nephrosclerosis, arterial hypertension, and cardiac hypertrophy.[2] Abundance of sodium in the diet causes these toxic effects of desoxycorticosterone to appear earlier and with increased severity.

a

b

Fig. 250. Addison's disease. *a*, cutaneous pigmentation; *b*, pigmentation of buccal mucosa, hands, and face. (*Courtesy of Dr. E. B. del Castillo.*)

[1] Desoxycorticosterone pellets can be implanted subcutaneously. They are gradually absorbed and have prolonged effects. Their activity is controlled by varying the dose of NaCl ingested.

[2] Selye, H., *Am. Heart J.*, **27**, 338, 1944.

Adrenal Insufficiency in Man

The principal forms of adrenal insufficiency that have been observed in man are (*a*) Addison's disease; (*b*) chronic adrenal insufficiency without pigmentation; (*c*) hypophyseal myxedema; (*d*) acute adrenal insufficiency.

The outstanding signs and symptoms of Addison's disease are asthenia, pigmentation of the skin and mucosae, low blood pressure, digestive disturbances (vomiting, diarrhea) and loss of body weight, but there are many other functional disturbances (Fig. 251).

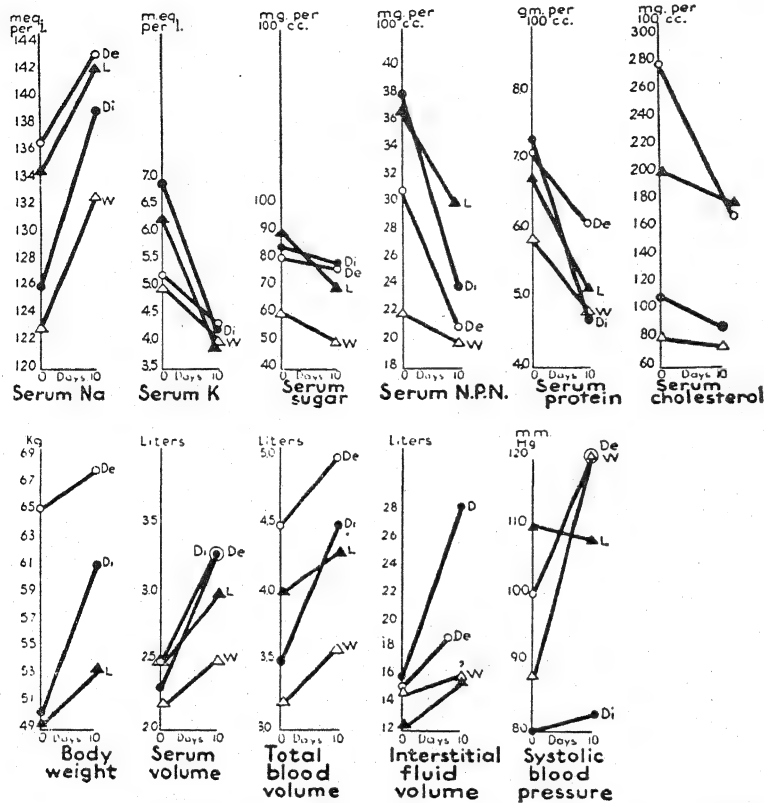



Fig. 251. Effect on patients with Addison's disease of 10 days' treatment with desoxycorticosterone derivative. (*Loeb, R. F., J. A. M. A., vol. 116, p. 2495, 1941.*)

Addison's disease. The disease is produced by extensive destructive lesions of the adrenals, most commonly due to tuberculosis of the glands (50 to 70 per cent of cases), or to simple atrophy of the cortex. The statistics of Addison's disease in certain countries show that 20 to 50 per cent of cases are due to simple atrophy of the adrenal cortex.

Three stages are observed in the course of the disease: (*a*) phase of compensation, in which the disturbances are slight or nonexistent; (*b*) phase of decompensation, in which all the effects of adrenal insufficiency are in evidence; (*c*) phase of acute adrenal insufficiency or adrenal crisis, which may lead to death by shock or circulatory collapse.

Asthenia consists of muscular weakness and physical and psychic fatigue. Pigmentation of the skin (melanodermia) is observed in 94 per cent of the cases. The skin takes on a bronze color and dark patches are seen in the mucosae.[1] The systolic blood pressure is usually below 90 mm., except in hypertensive subjects. The heart volume, as appreciated by radiography, is small. The loss of weight is due in part to anorexia.

Digestive symptoms consist of anorexia (which appears in the early stages and is marked and persistent), vomiting, and diarrhea; in some cases, there is abdominal pain.

The BMR is below normal in approximately one-

[1] These dark patches are sometimes seen in normal individuals with a dark skin.

half the cases; the body temperature is low. The following disturbances in carbohydrate metabolism are seen: (*a*) low fasting blood-sugar level (Porges); (*b*) hypersensitiveness to insulin;[1] (*c*) a slight increase in blood sugar after the ingestion of glucose, followed by marked secondary hypoglycemia. The RQ is usually high and rises considerably after the ingestion of glucose. There are no disturbances in salt metabolism during the compensation phase, but decompensation is preceded by loss of sodium in the urine, a fall in Na and Cl, and a rise in K in plasma; these changes are marked during the adrenal crisis.

Robinson, Power, and Kepler[2] have devised a diagnostic test for adrenal insufficiency. It is based on the fact that copious ingestion of water is followed by only slight diuresis and very little dilution of the urine, and that chlorides are excreted in abnormally large quantities, while urea and sodium are retained. Cutler, Power, and Wilder[3] have proposed another test which consists in restricting the intake of sodium and increasing the intake of potassium. This causes an increase in the severity of the urinary and plasma signs and general symptoms of adrenal insufficiency; acute insufficiency may be provoked. Frequently the lymphocyte and eosinophil count are high. Thorn's test[4] consists in the intravenous injection of 20 mg. of adrenocorticotrophin in the course of 8 hr.; an eosinophil count is made at the tenth hour, and the 24-hr. excretion of 17-ketosteroids is determined. Eosinopenia and an increase in 17-ketosteroid excretion are seen in normal subjects but not when the adrenal cortex does not function.

Sexual disturbances such as loss of libido and impotence in men and amenorrhea in women are frequently observed. The urinary elimination of 17-ketosteroids diminishes. Treatment with adrenal extracts or steroids and salt improves the patient's condition.

Hypophyseal myxedema. Certain patients with myxedema and a low BMR (−20 to −45 per cent) on receiving thyroid treatment enter into an adrenal crisis which may be fatal if it is not adequately treated. In cases of myxedema, adrenal function is usually more or less depressed.

Hypophyseal cachexia. Adrenal insufficiency is evident in some cases of Simmonds' disease.

[1] According to Marañon, insulin should never be given to patients with Addison's disease, because it may cause fatal hypoglycemia.

[2] ROBINSON, F. J., M. H. POWER, and E. J. KEPLER, *Proc. Staff Meet., Mayo Clin.*, **16**, 577, 1941.

[3] CUTLER, H. H., M. H. POWER, and R. M. WILDER, *J. A. M. A.*, **111**, 117, 1938.

[4] RENOLD, A. E., D. JENKINS, P. H. FORSHAM, and G. W. THORN, *J. Clin. Endocrinol.*, **12**, 763, 1952.

Adrenal crisis. Acute adrenal insufficiency is seen in the course of Addison's disease (adrenal crisis) and after operations for corticoadrenal tumors or adrenalectomy performed in cases of cancer of the adrenals or the prostate and malignant hypertension. Typical signs of adrenal insufficiency appear, but they can be controlled by treatment with desoxycorticosterone, cortisone, and NaCl.

In the course of chronic adrenal insufficiency several causes may provoke an adrenal crisis, *e.g.*, infections, surgical operations, violent exercise, exposure to cold, intravenous injection of glucose, saline purges, administration of thyroid preparations, morphine, or barbiturates. The patient falls into circulatory collapse, the blood pressure falls, and the pulse rate increases. The skin becomes cold and cyanotic. There are signs of dehydration and hemoconcentration. Plasma sodium and chloride fall, and urea and potassium rise. There are nausea, vomiting, and diarrhea. Asthenia is marked, mental reactions are sluggish, and sometimes the patient becomes delirious.

Acute adrenal insufficiency. This condition is sometimes called the *Sergent-Bernard syndrome.*[1] There are three different types, according to the nature of the symptoms: (*a*) the abdominal, pseudoperitoneal, or choleriform type; (*b*) the nervous or pseudomeningeal type; (*c*) the cardiovascular type.

The *Waterhouse-Friderichsen* syndrome is due to adrenal hemorrhage, provoked, in the majority of cases, by meningococcus. Its principal signs are cyanosis and edema of the face, red and hemorrhagic patches on the skin, oliguria or anuria, progressive peripheral circulatory collapse, and stupor or coma.

Treatment. Chronic adrenal insufficiency is treated by the following methods: (*a*) ingestion of 5 to 20 gm. daily of sodium chloride, reducing the intake of potassium from the usual 4 gm. to 2 gm. daily; (*b*) desoxycorticosterone (usually the acetate) injected subcutaneously (5 mg. daily) or by the implantation of pellets (125 to 500 mg.) associated with sodium chloride (3 gm. daily); (*c*) cortisone (25 mg. daily or more) associated with NaCl and desoxycorticosterone (0.5 to 1 mg. daily).

Corticoadrenal extracts are very efficacious but used less frequently, except in adrenal crises or when there is danger of a crisis (surgical operations, fever, etc., in patients with adrenal insufficiency). Desoxycorticosterone maintains life and the electrolyte and water balance, but care should be taken that hypo-

[1] SERGENT, E., *Presse méd.*, May 12, 1943; *Journées méd. Paris*, p. 528, 1937.

glycemia does not occur, as it may if the patient is not adequately fed. An excess of desoxycorticosterone may cause edema, hypertension, cardiac dilatation, and pulmonary edema, and the patient may die. The dose should be diminished as soon as the patient's condition improves or there is an increase in weight (1 lb. per day), or edema, or signs of cardiac dilatation.

Fig. 252. Corticoadrenal virilism. *a*, preoperative appearance of woman, unshaved; *b*, preoperative appearance, arms shaved; *c*, 2 years and 10 months after removal of tumor. (*Lukens, F. D. W., and J. D. Palmer, Endocrinol., vol. 26, p. 941, 1940.*)

Cortisone has a slower action; it is marked only after 5 to 10 days' treatment. The blood sugar and sugar tolerance return to normal, electrolyte balance is restored, the electroencephalogram becomes normal, etc.

In an adrenal crisis special care should be taken to maintain a normal blood-plasma volume and blood pressure and to see that hypoglycemia does not occur. Large amounts (1 to 1.5 liters) of NaCl solution (0.9 per cent) and glucose solution (5 per cent) are given, adrenal extract is injected subcutaneously or intravenously, desoxycorticosterone (20 mg.) is injected intramuscularly, and the patient is fed frequently with carbohydrates.

Adrenal extracts do not modify the melanodermia. The skin, however, becomes paler as a patient recovers from a crisis because of hydration of the tissues.

ADRENAL HYPERFUNCTION

Under this heading the following subjects will be discussed: (*a*) effects of treatment with adrenocorticotrophin or cortisone; (*b*) the general adaptation syndrome; (*c*) chronic adrenal hyperfunction in man.

Effects of adrenocorticotrophin and corticoadrenal hormones. Adrenocorticotrophin (ACTH) causes adrenal hypertrophy, hyperplasia, and hyperfunction. Its activity increases considerably if it is given in continuous injection. Its most typical effects (which are not seen if the adrenals have been removed or destroyed by disease) are a fall in blood eosinophils (80 to 100 per cent) and an increase in the urinary excretion of 17-ketosteroids and 11-oxysteroids. Both these are due to an increase in corticoadrenal secretion.

Prolonged treatment with cortisone in animals causes inhibition of the adrenals (atrophy and hypofunction). A similar effect has been observed in man; in these cases interruption of cortisone treatment is followed by signs of adrenal hypofunction: asthenia, decreased urinary excretion of 17-ketosteroids, and diminished response to ACTH.

Prolonged administration of ACTH or cortisone provokes the appearance of many signs of Cushing's disease (see pages 623–624); the principal disturbances are increase in appetite and body weight, rounding of the face, acne and hirsutism (signs of androgenic activity), and cutaneous striations.

ACTH frequently produces pigmentation in the skin. Some preparations contain the melanodispersing hormone, which can be separated from the corticotrophic activity (Li), although it has been maintained that they are secreted together (Sulman).

During treatment there are remarkable psychic changes; usually there is euphoria, but

in some cases the patient becomes depressed. Convulsions and mental disturbances have also been reported.

There is little or no change in blood pressure (Sprague). In patients with Addison's disease there is not much retention of sodium, contrasting in this with the effect of desoxycorticosterone. In subjects with functioning adrenals it may cause retention of Na and water which is later compensated. ACTH also causes transitory retention of Na, Cl, and water in subjects with intact adrenals. Both substances increase potassium excretion and may cause the K concentration in plasma to fall below the normal, with K deficiency in extracellular and intracellular fluids. In some cases there is alkalosis with low Cl and K concentration in blood (hypokalemic alkalosis), as may occur in patients with Cushing's disease.

In man the blood sugar usually remains normal; there is no glycosuria; sugar tolerance occasionally diminishes, but in most cases it is not affected; sensitiveness to insulin diminishes. In a few cases ACTH has provoked hyperglycemia and glycosuria, partly of renal origin (Conn). It increases the severity of the disease in diabetic subjects (Sprague, Thorn). Large doses may produce a temporary diabetes in the rat. Cortisone does not provoke diabetes in dogs with an intact normal pancreas, but after the pancreatic mass has been reduced surgically, cortisone may provoke transitory, corticoid, diabetes or permanent, metacorticoid, diabetes.

In the rat cortisone increases the capacity for muscular work, but in man it may cause muscular weakness, especially in cases of hypokalemia.

Large doses increase protein catabolism and the excretion of N, P, and K; effects which are not produced by small or medium doses. Animal experiments have shown an increase in protein catabolism and a decrease in protein anabolism.

Uric-acid excretion increases a few hours after injection of cortisone (Thorn), owing to its action on the kidney.

An optimum level of adrenal secretion is necessary for growth. Adrenalectomized animals either fail to grow or show retarded growth. An excess of cortisone, or of ACTH in animals with intact adrenals, also retards growth.

Storage and liver fat may increase. Ketonuria increases in diabetics and occasionally in normal subjects.

There is marked atrophy of lymphoid tissues in several species, but less in man. A transitory lympholytic effect has also been seen on malignant lymphatic tumors.

The blood picture shows neutrophilia, lymphopenia, and marked eosinopenia. In man lymphopenia is not so constant or so intense as in other species.

Inhibition and atrophy of the sexual organs have been observed in males; this effect is mediated by the hypophysis and does not occur in hypophysectomized animals. In women amenorrhea may occur, and in men the libido may diminish.

Thyroid function is more or less depressed in the course of treatment with ACTH or cortisone.

Large doses inhibit growth, osteogenesis, and chondrogenesis. The skin becomes thin, sebaceous glands are atrophied, and hair growth ceases. Hyperactivity of the mesenchyma in response to aggression by chemical or biological agents, and in allergy, is also inhibited. Cortisone diminishes inflammatory reaction and the development of fibroblasts and fibrous tissue; it retards healing of wounds and inhibits hyaluronidase, and in some cases antibody formation is subnormal. This decrease in defensive reactions favors the growth and dissemination of bacteria and other germs; thus, in patients suffering from tuberculosis the disease frequently increases in severity and has a tendency to spread. Inhibition of local inflammatory reaction may, however, be advantageous, *e.g.*, in some cases of allergy, or inflammatory processes of the eye or joints.

Treatment with ACTH or cortisone rapidly improves the symptoms of rheumatoid arthritis; in some cases within a few hours spectacular results are obtained,[1] Relapses, however, occur when the treatment is suspended, and occasionally even in the course of treatment.

Resistance to cold, hypoxia, fasting, histamine, etc., is increased by treatment with cortisone in adrenalectomized animals, and in a lesser degree in normal animals.

The general adaptation syndrome.[2] Selye has given this name to "the sum of all nonspecific systemic reactions of the body which ensue upon long exposure to stress." The reac-

[1] HENCH, P. S., *et al.*, *Ann. Rheumat. Dis.*, **8**, 90 and 97, 1949.

[2] SELYE, H., *J. Clin. Endocrinol.*, **6**, 117, 1946; "Textbook of Endocrinology," 2d ed., Acta Endocrinologica, Montreal, 1949; "Stress," Acta, Inc., Montreal, 1950.

tions are the same to all harmful agents, but they vary to some extent from one species to another.

The following circumstances provoke the appearance of the adaptation syndrome: (*a*) trauma (accidental, surgical, obstetrical); (*b*) violent muscular exercise; (*c*) infections; (*d*) hemorrhage; (*e*) cold; (*f*) fever; (*g*) anoxia; (*h*) burns; (*i*) poisons, or drugs and hormones in abnormally large doses; (*j*) bacteria; (*k*) radiation; (*l*) nervous shock; etc.

The syndrome develops in three stages: (*a*) the *alarm reaction* to sudden exposure to stimuli to which the organism is not adapted (this reaction has two phases, shock and countershock); (*b*) the *stage of resistance*, during which symptoms improve or disappear; (*e*) the *stage of exhaustion*, during which the symptoms reappear.

During the alarm reaction there are notable structural and functional changes in many organs, together with certain metabolic disturbances.

The functions of the hypophysis are rapidly and markedly disturbed. Adrenocorticotrophin secretion increases and causes corticoadrenal hypertrophy with a decrease in the ascorbic-acid, cholesterol, and fat content of the adrenal tissue. There is also an increase in the secretion of corticoadrenal hormones, which is responsible for some of the disturbances observed. These disturbances do not occur in adrenalectomized or hypophysectomized animals. They are the following: (*a*) atrophy of the thymus and lymphoid tissue, with a decrease in the blood lymphocyte count; (*b*) disturbances in water, salt, carbohydrate, and protein metabolism; (*c*) increase in the urinary elimination of corticoadrenal steroids. The increase in adrenocorticotrophin secretion has been demonstrated by the rapid decrease in the ascorbic-acid content of the adrenal, which does not occur in hypophysectomized animals.[1]

The increase in adrenocorticotrophin in man[2] causes a rapid decrease in the blood lymphocyte count, which has been observed in aviators flying to great heights, after the injection of adrenaline or insulin, etc. There is also an increase in the urinary elimination of "corticoids," *i.e.*, of

[1] LONG, C. N. H., *Federation Proc.*, **6**, 461, 1947; SAYERS, G., and M. A. SAYERS, *Recent Progress in Hormone Res.*, **2**, 81, 1948; Symposium on the Adrenals, *J. Endocrinol.*, **5**, No. 4, 1947.

[2] VENNING and BROWNE, *op. cit.*, FORBES, *et al.*, *op. cit.;* SHIPLEY, *et al.*, *op. cit.;* PINCUS, G., *Recent Progress in Hormone Res.*, **1**, 123, 1947.

substances with adrenal hormonal activity such as increasing the resistance of adrenalectomized rats to cold, or increasing the liver-glycogen content of adrenalectomized animals. These substances release formaldehyde when oxidized with periodic acid.[1] Another sign of corticoadrenal hyperactivity is an increase in the urinary excretion of 17-ketosteroids. Increases to 20 and 30 times the normal amount of urinary corticoadrenal steroids during approximately 2 days, followed by a decrease, have been observed in subjects submitted to surgical operations, flying to great heights, or suffering from burns, hemorrhage, anoxia, etc.

Adrenocorticotrophin secretion increases considerably and very rapidly. This provokes a decrease in the ascorbic acid,[2] cholesterol, and fat content of the adrenal, which becomes congested and later is hypertrophied. The secretion of corticoadrenal hormones is increased, and this increase is the cause of several signs which do not appear if the adrenals have been removed, *e.g.*, (*a*) atrophy of the thymus and lymphoid tissues, and lymphopenia in the blood; (*b*) marked decrease of blood eosinophils; (*c*) changes in water, salt, carbohydrate, and protein metabolism; (*d*) increase in the urinary excretion of 17-ketosteroids and corticosteroids.

In man accidental or surgical trauma, burns, hemorrhage, hypoxia, etc., are accompanied by the appearance of similar signs, *e.g.*, eosinopenia, and sometimes lymphopenia; increase in urinary corticosteroids, 17-ketosteroids, or 11-oxysteroids. In some cases increases up to twenty or thirty times the normal amount of urinary corticosteroids are observed for about two days, followed later by a decrease.

Chronic adrenal hyperfunction. There are several clinical forms of corticoadrenal hyperfunction (also called hypercorticalism), depending on age and sex of the patient and on which hormones are secreted in excess. In Cushing's disease metabolic disturbances predominate. Excess of androgenic hormones causes somatic or sexual precocity and, in adult women, virilization. Less frequently there is an excessive estrogenic activity which causes feminization in males.

[1] This reaction is usually attributed to "oxycorticosteroids" originated in the adrenal, but it is not specific for these substances.

[2] SAYERS, G., and M. A. SAYERS, *Recent Progress in Hormone Res.*, **2**, 81, 1948; *Physiol. Rev.*, **30**, 241, 1950.

Cushing found basophil-cell adenomas in the hypophysis, and attributed to hypophyseal basophilism the disturbances seen in the syndrome now called "Cushing's disease." The clinical signs and symptoms were seen to be those of corticoadrenal hyperfunction. Moreover, modifications in the basophil cells of the hypophysis are usually observed in cases of hypercorticalism (Crookes), and treatment with corticosteroids produces similar changes in the basophil cells. Therefore corticoadrenal hyperfunction is considered to be usually (perhaps always) the primary cause, and hypophyseal disturbances secondary to adrenal hyperfunction.

The most frequent signs and symptoms are the following: (*a*) *adiposity* of face, neck, and trunk, but not of the limbs; (*b*) *kyphosis*, due to osteoporosis of the cervicodorsal vertebrae (sometimes osteoporosis affects other bones, and frequently there is a negative calcium balance); (*c*) *arterial hypertension*, which usually develops in the later stages; (*d*) *diabetes*, or a diminished glucose tolerance, frequently observed; (*e*) *weakness*, asthenia, muscular atrophy, and a tendency to become tired easily; (*f*) *cutaneous signs*, such as purple atrophic striae and thinning of the skin, keratosis of the hair follicles, sometimes hypertrichosis, and a tendency to hemorrhages; (*g*) *sexual disturbances*, such as amenorrhea in women, loss of libido and impotence in men; (*h*) changes in *blood plasma*, *i.e.*, alkalosis with low Cl and K concentration; (*i*) changes in the *blood picture*, lymphopenia, eosinopenia and occasionally polycythemia; (*j*) *psychic* disturbances.

In a few cases of Cushing's disease irradiation of the hypophysis with x-rays has been followed by improvement. Subtotal adrenalectomy has given better results with considerable improvement in many cases.[1]

Adrenogenital syndromes are observed when there is hypersecretion of corticosteroids with sexual activity. When hypercorticalism develops during fetal life the child is a female pseudohermaphrodite with peniform clitoris and other signs of virilization. Boys affected in early childhood have a precocious puberty, *e.g.*, in a five-year-old boy the development of the penis and sexual hair may be that of a full-grown man. In girls heterosexual development (virilization) is observed: there are hypertrophy of the clitoris and abnormal hair growth (Fig. 252), the female sexual organs (ovaries, uterus, mammary glands) are not enlarged, and menstruation does not occur at an early age. Great muscular development and obesity in some cases give these children a herculean aspect. Girls are large and fat and take on a matronly appearance.

In adult females there is virilization and great development of the body hair (hirsutism). The hair on the head has the distribution and aspect of the male, there is a thick beard and mustache, and in many cases the whole body is covered with hair (Fig. 252). The skin is rough and thickened, and acne is often seen. The body build is that of a male, with broad shoulders and well-developed muscles. The voice is deep owing to the development of the larynx. There are many sexual disturbances: (*a*) menstruation is scarce or may be absent (amenorrhea); (*b*) sterility is observed in 50 per cent of the cases; (*c*) the mammary glands are atrophic; (*d*) anomalies in the libido occur frequently. With age other symptoms appear, such as obesity; purple striae in the skin; high blood pressure; osteoporosis, limited to the cervicothoracic vertebrae (kyphosis) or more widely spread; latent or manifest diabetes.[1] Often there is a negative nitrogen balance, accompanied by muscular weakness.

Hypercorticalism is observed more frequently in women than in men, but a few cases of feminization have been reported in men who have testicular atrophy and development of the mammary gland.

Hypertrichosis and the abnormal development of male or female sexual characters have been attributed to androgenic and estrogenic steroids secreted by the hyperactive adrenal cortex.

In many cases surgical extirpation of the adrenal tumor was followed by the total disappearance of the signs and symptoms of hypercorticalism. The tumor may have caused inhibition of the normal adrenal tissue, and when it is removed frequently acute corticoadrenal insufficiency develops. Precautions should therefore be taken during the preoperative and postoperative stages to avoid an adrenal crisis.

[1] Priestley, J. T., *et al.*, *Ann. Surg.*, **134**, 464, 1951; Kuppermann, H. S., *et al.*, *J. Clin. Endocrinol.*, **11**, 774, 1951.

[1] The Achard-Thiers syndrome (diabetes in bearded women) is one type of hypercorticalism.

The determination of the urinary excretion of steroids may give valuable information on the functional state of the adrenal cortex. In cases of Cushing's disease the excretion of 17-ketosteroids is usually normal or only slightly above normal; the excretion of glycocorticoids (11-oxysteroids) is increased (Venning and Browne). In adrenogenital syndromes and in hirsutism, 17-ketosteroid elimination is increased and that of 11-oxysteroids remains normal. In Addison's disease, after adrenalectomy, in hypophyseal insufficiency and in anorexia nervosa, excretion of 17-ketosteroids and 11-oxysteroids diminishes considerably. In acromegaly it is usually increased. In adrenal hyperplasia and hyperfunctioning tumors of the adrenal cortex urinary 17-ketosteroids are considerably increased; they return to normal levels after the tumor has been removed. Trauma and violent exercise increase urinary corticosteroids; an increase is also observed during the last weeks of pregnancy.

Abnormal hyperactivity of the adrenal medulla. In patients with chromophil-cell tumors, paroxysmic or permanent hypertension is observed. These tumors contain and secrete adrenaline and noradrenaline in variable proportions; frequently noradrenaline predominates. Hypertensive crises occur spontaneously and may be provoked by intravenous injection of 0.025 mg. of histamine (Kvale and Roth). Permanent hypertension due to chromaphil-cell tumors is reduced and hypertensive paroxysms are prevented by the injection of adrenolytic drugs (benzodioxane, regitine, etc.). Extirpation of the tumors lowers the blood pressure, if it was above normal, and suppresses the hypertensive crises.

BIBLIOGRAPHY

The Adrenal Cortex, *Ann. New York Acad. Sc.*, **50**, 678, 1949.

Conf. on the Adrenal Cortex, Trans. (Josiah Macy, Jr., Foundation Publications), New York, 1949 to 1952.

Dorfman, R. I., The Bioassay of Adrenal Cortical Steroids, *Ann. New York Acad. Sc.*, **50**, 556, 1949.

Fourman, P., *et al.*, Effects of 17-hydroxycorticosterone (Compound F) in man, *J. Clin. Investigation*, **29**, 1462, 1950.

Gaunt, R., The Adrenal Cortex in Salt and Water Metabolism, *Physiol. Rev.*, **29**, 281, 1949; *Recent Progress in Hormone Res.*, **6**, 247, 1951.

Hartman, F. A., and K. A. Brownell, "The Adrenal Gland," Lea & Febiger, Philadelphia, 1949.

Hechter, O., A. Zaffaroni, R. P. Jacobsen, H. Levy, R. W. Jearloz, V. Schenker, and G. Pincus, The Nature and Biogenesis of the Adrenal Secretory Products, *Recent Progress in Hormone Res.*, **6**, 215, 1951.

Hume, D. M., G. W. Harris, J. de Groot, D. Jacobsohn, C. Fortier, C. N. H. Long, and J R. Brobeck, Control of ACTH Secretion by the Pituitary Gland, "Ciba Foundation Colloquia on Endocrinology," Vol. 4, 1952, pp. 87–145.

Ingle, D. J., Some Studies on the Role of the Adrenal Cortex in Organic Metabolism, *Ann. New York Acad. Sc.*, **50**, 576, 1949.

Julian, P. L., Chemistry of the Adrenal Cortex Steroids, *Recent Progress in Hormone Res.*, **6**, 195, 1951.

Kendall, E. C., The Chemistry and Partial Synthesis of Adrenal Steroids, *Ann. New York Acad. Sc.*, **50**, 540, 1949.

Kendall, E. C., The Influence of the Adrenal Cortex on the Metabolism of Water and Electrolytes, *Vitamins & Hormones*, **6**, 277, 1948.

Kepler, E. J., and W. Locke, Hiperfunción Suprarrenal Aguda, in "Tratado de Endocrinología Clínica," Vol. 2, G. Kraft, Buenos Aires, 1950, p. 185.

Kepler, E. J., R. G. Sprague, H. L. Mason, and M. H. Power, The Pathologic Physiology of Adrenal Tumors and Cushing's Syndrome, *Recent Progress in Hormone Res.*, **2**, 345, 1948; *Ann. New York Acad. Sc.*, **50**, 657, 1949.

Li, C. H., and H. M. Evans, The Properties of Growth and Adrenocorticotrophic Hormones, *Vitamins & Hormones*, **5**, 198, 1948.

Loeb, R. F., The Adrenal Cortex and Electrolyte Behavior, *J. A. M. A.*, **116**, 2495, 1941; *Harvey Lect.*, **37**, 100, 1942.

Long, C. N. H., A Discussion of the Mechanism of Action of Adrenal Cortical Hormones on Carbohydrate and Protein Metabolism, *Endocrinology*, **30**, 870, 1942.

Marañon, G., Insuficiencia Suprarrenal, in "Tratado de Endocrinología Clínica," Vol. 2, G. Kraft, Buenos Aires, 1950, p. 123.

Morató-Manaro, J., Nociones Mortológicas y Fisiología de las Suprarrenales, in "Tratado de Endocrinología Clínica," Vol. 2, G. Kraft, Buenos Aires, 1950, p. 45.

Mote, J. R., "Proceedings of the Clinical ACTH Conferences," Blakiston, Philadelphia, 1950–1951.

Noble, R. L., Physiology of the Adrenal Cortex, in "The Hormones," Vol. 2, Academic Press, New York, 1950, p. 65.

Reichstein, F., and C. W. Shoppee, The Hormones of the Adrenal Cortex, *Vitamins & Hormones*, **1**, 345, 1943.

Sayers, G., The Adrenal Cortex and Homeostasis, *Physiol. Rev.*, **30**, 241, 1950.

Sayers, G., and M. A. Sayers, The Pituitary Adrenal System, *Recent Progress in Hormone Res.*, **2**, 81, 1948; *Ann. New York Acad. Sc.*, **50**, 522, 1949.

Schweiz. Akad. Med. Wissensch., Symposium on the Influence of the Hypophysis and Adrenal Cortex on Biological Reactions, Schwabe, Zurich, 1951.

Selye, H., "Stress," Acta, Inc., Montreal, 1950–1951.

Selye, H., "Textbook of Endocrinology," 2d ed., Acta Endocrinologica, University of Montreal, Montreal, 1949.

Shoppee, C. W., Recent Advances in the Chemistry of the Adrenal Cortex, *J. Endocrinol.*, **5**, 27, 1947.

Soffer, L. J., "Diseases of the Adrenals," 2d ed., Lea & Febiger, Philadelphia, 1948.

Sprague, R. G., Effects of Cortisone and ACTH, *Vitamins & Hormones*, **9**, 263, 1951.

Sprague, R. G., *et al.*, Observations on the Physiologic Effects of Cortisone and ACTH in Man, *Arch. Int. Med.*, **85**, 199, 1950.

Thorn, G. W., "The Diagnosis and Treatment of Adrenal Insufficiency," *American Lecture Series*, Publ. No. 29, Charles C Thomas, Springfield, Ill., 1949.

Thorn, G. W., *et al.*, Advances in the Diagnosis and Treatment of Adrenal Insufficiency, *Am. J. Med.*, **10**, 595, 1951.

Venning, E., and J. S. L. Browne, Urinary Excretion of Adrenal Cortical Steroids, *Ann. New York Acad. Sc.*, **50**, 627, 1949.

CHAPTER 55

The Parathyroid Glands

THE PARATHYROIDS ARE small endocrine glands of great physiologic importance. Their removal or insufficiency causes serious disturbances, which may be fatal if adequate treatment is not given. Parathyroid secretion controls homeostasis of phosphorus and calcium; it maintains a normal blood-calcium level and raises it if it falls below the normal. It regulates the deposition and mobilization of calcium salts in the bones and teeth; therefore it modifies the structure and consistency of the skeleton. It controls calcium and phosphorus metabolism and the renal excretion of calcium and phosphate. Through its effect on calcium metabolism and blood calcium, it controls the ionic equilibrium of body fluids and is thus an important factor in the maintenance of the normal excitability of the tissues and the permeability of cell membranes.

Historical note. The parathyroids were seen by Remak in 1855, but they were not fully described until Sandström published a very complete paper on them in 1880. Gley (1891) attributed to them some of the results observed after thyroidectomy and considered them undeveloped thyroid tissue. Moussu (1896) and Vassale and Generali (1896) demonstrated the differences in the functions of the thyroid and parathyroids. Before this work was published most so-called "thyroidectomies" were really thyroparathyroidectomies, and the "acute symptoms" of thyroidectomy, *e.g.*, tetany, were really due to parathyroidectomy. It is now fully understood that the thyroid and parathyroids have different structure, embryologic development, and functions.

Physiologic anatomy. There are usually two parathyroids on each side. One, situated on or near the thyroid, is the upper, or external, parathyroid. It is also known as "parathyroid IV" because it develops from the endoderm of the fourth branchial cleft. The lower, or internal, parathyroid is usually found within the thyroid gland. It is also called "parathyroid III" because it develops from the endoderm of the third branchial cleft. In the rabbit the external parathyroids are situated some distance from the thyroid; in the dog and cat they are placed on the upper pole of the thyroid. When thyroidectomy is performed in the rabbit the external parathyroids are not damaged, but thyroidectomy in the dog or cat must be performed with great care if the external parathyroids are to remain intact. Anomalies in position and accessory parathyroids occur frequently. The principal parathyroids in man have the shape of oval disks measuring 6 to 7 mm. by 3 to 4 mm. by 1 to 2 mm. All the parathyroid tissue together weighs approximately 140 mg.

There are two main types of parathyroid cells:

1. The principal, or chromophobe, cells are the most numerous. They have a large nucleus and no granules.
2. The eosinophil cells are less numerous, and larger than the former. They contain eosinophil granules. Eosinophil cells are not seen in man in early childhood; they appear about the tenth year of age. They are not found in the dog and other species.

PARATHYROID INSUFFICIENCY

Total parathyroidectomy can be performed without removing the thyroid, but usually thyroparathyroidectomy is performed when all the parathyroid tissue is to be removed, because the internal parathyroids are not always easily found. In certain cases there are accessory parathyroids situated some distance from the thyroid; they are often found included in the thymus.

Total parathyroidectomy is followed by tetany (parathyroid tetany). The cat and dog have been used most often for the experimental

study of parathyroid tetany, but it has been observed in many other species, *e.g.*, rats and other mammals, reptiles, and amphibians, and it has been observed in man after thyroidectomy performed in cases of goiter. Tetany appearing spontaneously has also been reported.

FIG. 253. Chronic hypoparathyroidism. Carpospasm provoked by compression of the arm by a pneumatic cuff (Trousseau's sign).

The name "tetany" was given by Corvisart in 1852 to convulsive conditions similar to those provoked by tetanic infection of wounds and certain drugs, such as strychnine. The different types of tetany were described in Chap. 45 when dealing with calcium metabolism.

There are three types of parathyroid insufficiency: (*a*) acute; (*b*) chronic; (*c*) latent. Total parathyroidectomy usually provokes acute tetany in the dog and cat, which ends in death within a few days in the majority of the animals. In man chronic tetany is more commonly observed.

Neuromuscular hyperexcitability. Hyperexcitability is a typical sign in all forms of tetany. It is the cause of the convulsions of acute tetany, but it is present and can be revealed by adequate tests in latent tetany. The following tests are those most commonly used: (*a*) if the arm is constricted by an armlet and the circulation is obstructed, there is a spasm of the muscles of the forearm and hand (carpospasm) which causes the hand to adopt a typical posture (Fig. 253) (Trousseau's sign); (*b*) tapping over the facial nerve where it is near the surface (in front of the ear) causes a spasm of the facial muscles on that side (Chvostek's sign); (*c*) the galvanic threshold of nerves and muscles is below normal (Erb's sign); (*d*) the rheobase is lower and chronaxie higher than in normal subjects (see Chap. 66); (*e*) hyperpnea of a few minutes duration provokes muscular twitchings and convulsions in subjects with latent tetany, owing to alkalosis.

The signs of tetany are mainly involuntary muscular contractions. In mild forms only twitchings or tremor are observed; in the severer forms, sometimes called spasmophilia, the following signs occur: (*a*) carpopedal spasm, *i.e.*, spasm of the hands and feet, in which

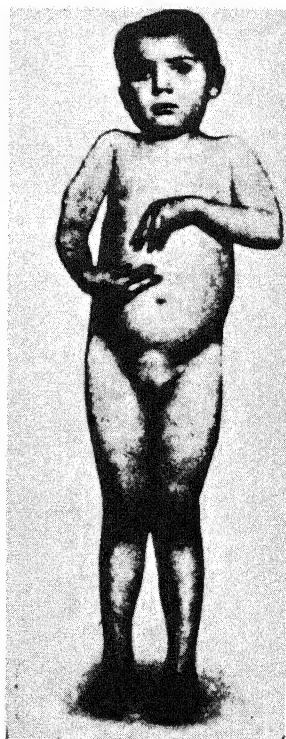



FIG. 254. Tetany. Carpopedal spasm. (*According to Falta, in Biedl, "Innere Sekretion," 4th ed., Urban & Schwarzenberg, Vienna, 1922.*)

the hands are flexed at the wrist, the fingers flexed at the metacarpophalangeal joint, extended at the interphalangeal joints, and brought together by adduction (the so-called "accoucheur's hand," Figs. 253 and 254); (*b*) laryngeal spasm, causing inspiratory stridor;

(*c*) a typical sardonic risus and a pained expression of the face; (*d*) spasms of other muscles and sometimes of the whole body (convulsions); (*e*) less frequently, spasms in smooth muscles, *e.g.*, the iris, esophagus, stomach, intestines, and bladder (visceral tetany); (*f*) changes in the electrocardiogram, which disappear with calcium treatment; these have been reported in the dog (Segura, 1937) and in man (Gotta and Pinto, 1941). The spasms are usually painful.

Spastic rigidity and convulsions in the form of epileptoid crisis are sometimes observed in man. The electroencephalogram shows signs of cortical hyperactivity in some cases, but not in all; they are related to the low blood-calcium level.[1]

In dogs and cats, contracture of the masticatory muscles keeps the jaws clamped. There is rigidity of the spinal muscles, and the limbs are spastically extended. Tremor and muscular twitchings increase gradually and merge into crises of generalized convulsions, separated by intervals of hypotonia and paresia. Muscular hyperactivity causes a rise in rectal temperature and thermal polypnea with salivation and tachycardia. Hepatic and renal lesions with albuminuria have been reported. The majority of untreated animals die in 4 to 7 days.

Spasticity and convulsions are due to hyperexcitability of the nerve centers. Extirpation of the cerebral cortex or decerebration (section of the brain stem) increase these symptoms. Spasticity is due to hyperactivity of the reflexes that produce normal muscle tone; it can be suppressed by section of the dorsal roots or the ventral roots, or by destruction of the spinal cord. Denervation of a muscle group suppresses spasticity and convulsions in that group, but fine fibrillar contractions may still be observed.

Disturbances in calcium and phosphorus metabolism. There are four typical signs of parathyroid insufficiency: (*a*) hypocalcemia; (*b*) hyperphosphatemia; (*c*) hypocalciuria; and (*d*) hyperphosphaturia. Blood calcium falls after parathyroidectomy from the normal level (9.5 to 11 mg. per cent) to 7 mg. per cent and in some cases 5 mg. per cent (McCallum and Voegtlin). Calcium ion concentration also decreases. The fall in blood calcium causes the excretion of calcium in the urine to diminish; calciuria ceases when the blood calcium is below 6 mg. per cent. The calcium in the feces is not significantly changed, but total calcium excretion diminishes owing to the decrease in urinary calcium elimination.

[1] Gotta, H., and J. B. Odoriz, *J. Clin. Endocrinol.*, **8**, 674, 1948.

Blood phosphate increases to 6 or 8 mg. per cent and in some cases to 12 mg. per cent. Urinary excretion of phosphate diminishes, but there is little or no change in the excretion of phosphorus in the feces. Alkaline phosphatase in blood increases.

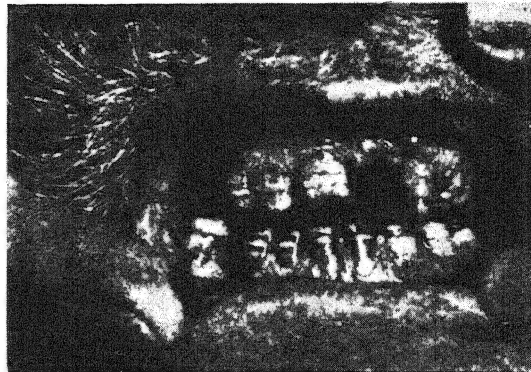



Fig. 255. Effect of parathyroid insufficiency on teeth. Erosion and discoloration of teeth in a patient with tetany. (*According to Phelps, in Biedl, "Innere Sekretion," 4th ed., Urban & Schwarzenberg, Vienna, 1922.*)

Other symptoms. Chronic hypoparathyroidism is accompanied by trophic disturbances, among which the following can be mentioned:

1. *Cataract* (opacity of the eye lens) is gradually established in prolonged hypocalcemia.
2. *Dental lesions* (Fig. 255) are marked in infantile tetany, less marked if the onset of the disease does not occur until the eighth to thirteenth year; they are not observed in adults. The rat is particularly appropriate for the experimental study of these lesions, because of the continuous growth of the incisors (Erdheim, Erausquin). The teeth take on a dull white aspect, and later brown stains appear (Fig. 256). They are easily deformed and fractured. The enamel is hypoplastic. The odontoblasts are irregularly placed and show signs of degeneration. The dentine is poorly calcified (areolar or globular dentine), and there is a large pulp cavity. According to Erdheim, if a parathyroid is grafted normal dentine is again formed.
3. *Bone lesions* are not prominent, but in young animals ossification is delayed, and consolidation of fractures is retarded in adults.

4. *Deficient development of the hair and nails* is observed. The nails are easily broken, and the hair falls.
5. *Cutaneous disturbances*, such as erosions and ulcerations, are sometimes observed.

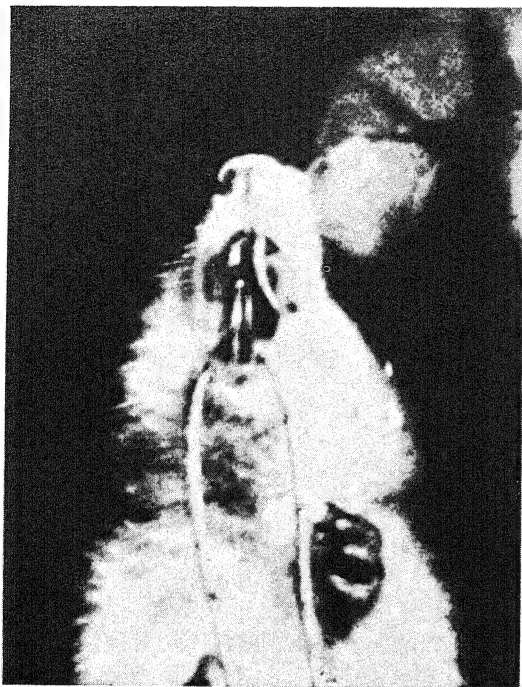



Fig. 256. Teeth of parathyroidectomized rat. Discoloration; fractured right upper incisor; excessive growth of left upper incisor.

Chronic and latent hypoparathyroidism. In some animals parathyroidectomy is not followed by acute tetany. The disease has a chronic course; the animals lose weight, trophic disturbances, such as cataract, appear after a time, and if the animals are young, growth is retarded. Calcemia is low in these animals, but spasticity is not observed, or may occur intermittently. In the so-called "latent tetany" there is persistent hypocalcemia, and sometimes trophic disturbances, but the usual signs of tetany (spasticity, muscular twitchings, convulsions) occur only in certain circumstances which provoke in these animals attacks of acute tetany.

Factors that modify the course of tetany. The severity of acute experimental tetany is increased, and latent tetany is made manifest, in the following circumstances: (*a*) a meat diet; (*b*) severe exercise; (*c*) estrus or pregnancy; (*d*) a high external temperature; (*e*) nervous or emotional stimulation; (*f*) injection of decalcifying agents (phosphates, etc.). Factors that diminish the severity of tetany usually increase the blood-calcium level and decrease excitability; the following may be mentioned: (*a*) injection of large quantities of Ringer's solution, which contains calcium; (*b*) a milk diet, which has a high calcium content; (*c*) administration of calcium, magnesium, or strontium salts; (*d*) vitamin D or cod-liver oil, which increase blood calcium (Fig. 257); (*e*) hemorrhage; (*f*) depressants of nervous excitability. Parathyroidectomized dogs can be kept alive and free from tetany by giving them 5 gm. of calcium lactate or gluconate twice daily. After 2 or 3 months of this treatment, it can be discontinued in many of the animals, which remain in latent tetany without symptoms in spite of the low blood calcium (Frouin, Luckhardt); apparently the tissues have become adapted to this low calcium level.

The mechanism of tetany. Several theories have been put forward to explain the mechanism of tetany. The principal ones postulate (*a*) the accumulation of a toxic substance in the organism (toxic theory); (*b*) a disturbance in the acid-base equilibrium; (*c*) a disturbance in calcium metabolism.

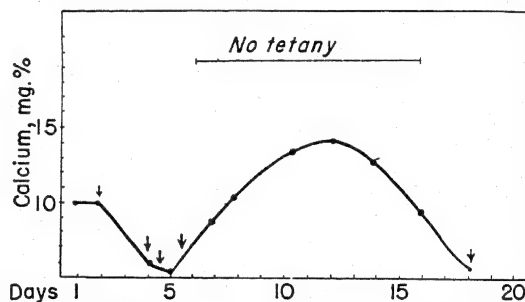



Fig. 257. Effect of calcium and vitamin D_2 on blood calcium and tetany. Arrows from left to right: 1, thyroparathyroidectomy; 2, tetany; 3 and 4, 4 doses of calcium gluconate (5 gm. twice daily for 2 days—20 gm. in all) and 40 mg. vitamin D_2 in one dose; 5, tetany. Ordinate, calcium concentration in plasma, in milligrams per cent; abscissa, time in days.

The toxic theory was the first to be proposed, and much work was done in the search for the toxic substance. Xanthine, histamine, carbamates, guanidine, and other compounds have been considered the specific poison that accumulates after parathyroidectomy and causes tetany. These substances provoke effects that are similar to, but not identical with, those of

tetany due to parathyroid insufficiency. They do not produce hypocalcemia, and the disturbances they cause are not controlled by calcium treatment or parathyroid extracts. It was frequently maintained that the poison producing tetany was absorbed from the intestine, but Billi has demonstrated that tetany occurs after parathyroidectomy in animals in which the stomach and intestines have been completely extirpated.

Alkalosis produced by the ingestion of an excess of alkali (*e.g.*, sodium bicarbonate) or by noncompensated loss of acid (*e.g.*, repeated vomiting with loss of HCl) causes tetany. Moreover acidosis, *e.g.*, that produced by the ingestion of ammonium chloride, improves the condition of animals or patients in parathyroid tetany because it favors the ionization of calcium in the body fluids. However, there is no alkalosis in tetany of parathyroid insufficiency; on the contrary, the convulsive periods cause a slight transitory compensated acidosis due to muscular hyperactivity. There is therefore no disturbance of acid-base equilibrium in parathyroid insufficiency capable of provoking tetany.

There are marked disturbances in calcium metabolism in parathyroid insufficiency, such as hypocalcemia, hypocalciuria, and lesions in the bones and teeth due to decalcification. Ingestion of a sufficient amount of calcium prevents or cures hypocalcemia and tetany, and all measures that are efficacious in the treatment of parathyroid tetany cause a rise in blood calcium. Thus parathyroid grafts or injections of parathyroid extracts increase blood-calcium concentration and control tetany.

The intimate mechanism that is the cause of hypocalcemia in parathyroid insufficiency is not known. At one time it was attributed to loss of calcium in the urine, but hypocalcemia occurs after parathyroidectomy in nephrectomized dogs, and parathyroid extracts cause the same rise in blood calcium in these animals as in controls with intact kidneys.

The decrease in calcium ion that occurs in parathyroid insufficiency disturbs the ionic equilibrium in body fluids, *i.e.*, the ratio of monovalent ions to bivalent ions, $\frac{(Na^+) + (K^+)}{(Ca^{++}) + (Mg^{++})}$, which Loeb showed was an important factor in neuromuscular excitability. A decrease in calcium concentration causes hyperexcitability of the nervous centers, which provokes spasticity (hypertonicity), tremor, muscular twitchings, and eventually convulsions, *i.e.*, tetany.

Parathyroidectomy causes a selective appetite for calcium. Parathyroidectomized rats that are allowed to choose between tap water and calcium solution drink large amounts of the latter. If these animals are treated with dihydrotachysterol, which increases the blood calcium, the consumption of calcium solution diminishes to the preoperative level. On the other hand, if parathyroidectomized rats are offered a phosphate solution, they drink less of it than normal controls (Richter).

Several objections have been made to the calcium-disturbance theory of tetany, among which the following may be mentioned: (*a*) hypocalcemia without tetany has been observed in cases of chronic nephritis, in animals fed on diets with an insufficient amount of calcium, and in parathyroidectomized dogs in which calcium treatment is discontinued after having been given during several weeks; (*b*) tetany can be improved without increasing the blood calcium by treatment with ammonium chloride or depressants of the central nervous system. These exceptional cases are usually interpreted as due to habituation of the tissues to a low blood calcium, or to the fact that ionic calcium is kept at a normal level in spite of the decrease in total calcium; thus ammonium chloride increases calcium ionization without raising the blood-calcium concentration.

Treatment of parathyroid insufficiency. Hypoparathyroidism is treated by raising the blood calcium or administering parathyroid hormone. Calcium treatment prevents death from acute tetany in dogs and carries them over to a condition of latent tetany. Calcium therapy is also the only method of rapidly relieving an acute attack of tetany, but it should be given intravenously only in cases of extreme emergency, because unless care is taken to inject it at a very slow rate, acute hypercalcemia may be provoked. Moreover most calcium salts are irritants (calcium gluconate is one of the least caustic) and cause damage to the vein (phlebitis). Usually 20 to 30 gm. of calcium lactate or gluconate is given daily by mouth. In this way it does not act so rapidly as when it is injected intravenously, but its effects last longer. Nevertheless the total daily amount should be given in several doses to obtain a more prolonged and smooth action. Blood calcium can be main-

tained stable at a high level by treatment with vitamin D or dihydrotachysterol, and patients with hypoparathyroidism can thus be kept free from tetany. Vitamin D_2 must be given in doses of 10 to 30 mg. per day (400,000 to 1,200,000 IU) until the blood calcium has reached a normal level; afterward the dose is reduced to 3 to 5 mg. per day (120,000 to 200,000 IU). With adequate supervision this treatment can be kept up for a long time. Dihydrotachysterol, introduced by Holtz under the name of AT 10 (antitetany preparation No. 10), is given in doses of 6 mg. daily to raise the blood-calcium level; afterward 1 to 3 mg. daily is enough to maintain a normal blood calcium. Large doses are toxic, and the patients should be carefully watched for signs of hypercalcemia. The diet should always contain an adequate amount of calcium, and it is advisable to reduce the phosphorus content to 0.5 or 0.6 mg. per day.

Active parathyroid extracts were prepared for the first time by Collip in 1925, under the name of "parathormone," but so far the parathyroid hormone has not been obtained pure.[1] This extract has the following effects: (*a*) it prevents or cures tetany in the dog and man; (*b*) it raises the blood-calcium level after an interval of 4 hr., reaching a maximum in 18 hr.; (*c*) it lowers the plasma phosphate; (*d*) it increases the urinary excretion of calcium and phosphate; (*e*) it mobilizes calcium stored in the bones. The effects of parathyroid extracts diminish gradually after a time and the dose must be raised to obtain constant results. Finally very large doses must be given to counteract the results of habituation.

Parathyroid glands have been grafted successfully in the rat (Cristiani, Erdheim) and in the dog (Halsted; Rojas and Manfredi). Successful grafts have also been obtained in human cases.[2] The practical importance of this method of treatment is not great, because it is not easy to obtain fresh healthy parathyroid tissue for grafting, and large numbers of the grafts are reabsorbed.

THE REGULATION OF PARATHYROID SECRETION

The internal secretion of the parathyroid glands maintains blood calcium at a normal

[1] Ross, W. F., and R. T. Wood, *J. Biol. Chem.*, **146**, 59, 1942.

[2] Manfredi, J. T., "Injertos Paratiroideos," J. Vicenti, Buenos Aires, 1944.

level and raises it if it has fallen below this level. If blood calcium is decreased by the injection of fluoride or oxalate, it returns to the normal concentration after a time only if the parathyroids are intact; in parathyroidectomized animals blood calcium remains low.[1] On the other hand, the blood-calcium curve following intravenous injection of calcium is not modified by parathyroidectomy. Hypocalcemia stimulates parathyroid secretion; this is demonstrated by the following facts: (*a*) a parathyroid graft made by vascular anastomosis restores the normal blood-calcium level within a few hours in parathyroidectomized dogs, but the blood-calcium concentration of normal animals is not modified by a parathyroid graft;[2] (*b*) if blood from an animal in hypocalcemia is perfused through a parathyroid gland and is then injected into another animal, it provokes hypercalcemia in the latter, but this is not observed if the blood perfused through the parathyroid has a normal calcium concentration (Patt, Wallerstein, and Luckhardt, 1942). Parathyroid secretion therefore controls the concentration of calcium in the blood and maintains it at a normal level. Reciprocally the blood-calcium level controls parathyroid secretion, so that the amount of hormone secreted is sufficient to maintain a constant normal blood-calcium concentration.

HYPERPARATHYROIDISM

Acute hyperparathyroidism. This is provoked by the administration of excessive doses of parathyroid extract. It has been observed and well studied in animals, and also in a few cases of patients treated with parathyroid hormone. Blood calcium rises to 18 and even to 20 mg. per cent. Calcium is deposited in several tissues (lung, kidney, gastric and intestinal mucosa) and causes damage there. At the same time signs of calcium intoxication appear: loss of appetite, depression, muscular weakness, polyuria, vomiting, diarrhea, and dehydration. In a later stage the blood calcium falls slightly (Fig. 258), when renal lesions cause an increase in the nonprotein nitrogen of the blood. Blood phosphate is low in the first period and rises later when the kidney is damaged. In advanced calcium intoxication there are muscular relaxation (loss of tone), somnolence, coma, and eventu-

[1] Gerschman, R., *Rev. Soc. argent. de biol.*, **6**, 25, 1930.

[2] Lewis, J. T., and R. Gerschman, *Rev. Soc. argent. de biol.*, **5**, 774, 1929.

ally death. In several species a very large dose of parathormone provokes acute decalcification of bones. Prolonged treatment with large doses produces osteoporosis and osteitis fibrosa. According to Selye, the parathyroid hormone acts on the osteoblasts; a large dose converts these cells first into osteoclasts, which reabsorb bone (calcium, phosphate, and organic matrix), and later into osteoblasts which form bone.

Chronic hyperparathyroidism. In 1891 von Recklinghausen described a disease of the bones to which the name "osteitis fibrosa cystica of von Recklinghausen" was given. Later hyperplasia or tumors of the parathyroids, which were thought to be a defensive response to the disturbance in calcium metabolism, were found in these patients. Schlagenhaufer suggested that they were the cause of the disease, and Mandl (1925) performed the first extirpation of the parathyroids in these cases, observing a remarkable improvement in the patient's condition. Later this fact was confirmed in many cases and corroborated by the experimental production of osteitis fibrosa cystica in animals by prolonged treatment with large doses of parathyroid extract.

In hyperparathyroidism there is *hypertrophy or hyperplasia of the parathyroids or a tumor* (adenoma or malignant growth) in one of the glands together with disturbances in calcium and phosphorus metabolism, lesions in the bones and kidney, and general symptoms.

Disturbances in calcium and phosphorus metabolism include hypercalcemia (12 to 15, exceptionally 20, mg. per cent). Blood phosphate falls below 3 mg. per cent, frequently to 1 and 2 mg. per cent, except in the late stages when renal lesions cause retention of phosphate. Urinary excretion of calcium and phosphate is abnormally high; normal amounts of calcium and phosphate are eliminated in the feces, but the calcium balance is often negative, owing to hypercalciuria. Plasma phosphatase also increases.

Bones. Calcium is mobilized and removed from the bones, which become decalcified and painful. The bones lose consistency and are deformed by the weight of the body. Thus the vertebrae are flattened, the spine is abnormally curved (kyphosis and scoliosis), and the subject's height decreases. The bones of the limbs show abnormal curvatures, and fractures sometimes occur spontaneously. Radiographs show general decalcification of the bones, but certain areas are more intensely decalcified, especially in the cranium, spine, and pelvis. Tumors and cysts are also formed, more often in the long bones of the limbs than elsewhere. In about one-third of the cases, in which the calcium balance is not negative, there are no lesions in the bones.

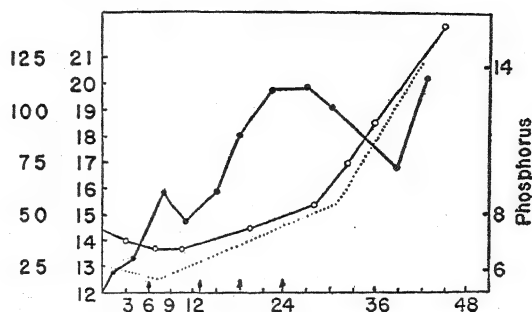



FIG. 258. Effect of parathyroid overdosage on blood calcium (dark line), phosphorus (light line), and nonprotein nitrogen (dotted line). At each arrow 25 units of parathormone was injected. Abscissa, time in hours; ordinates, concentration of nonprotein nitrogen, calcium, and phosphorus in the blood, in milligrams per cent. (*Collip, J. B., Medicine, vol. 5, p. 1, 1926.*)

Kidneys. Renal lesions are seen in 92 per cent of the cases; they are secondary to the disturbances in calcium metabolism. Polyuria often occurs; sometimes it is as extreme as in patients with diabetes insipidus, and provokes intense thirst and polydipsia. Calcium is deposited in the cells of the renal tubes, and there is renal insufficiency. Frequently calcium phosphate or oxalate precipitates in the urine (urinary "sand") or stones are formed in the collecting tubes or the renal pelvis and may provoke ureteral obstruction (renal colic).

General symptoms. Calcium may precipitate in the intima of the arteries, in the gastric mucosa, bronchi, and other tissues. Hypercalcemia produces a decrease in muscle tone, muscular weakness (asthenia), neuromuscular hypoexcitability, and difficulty in performing movements. In more serious cases there is loss of appetite or vomiting, loss of weight, and anemia, and constipation occurs frequently.

Treatment of hyperparathyroidism. The diet should contain a large amount of calcium and a low phosphorus content to prevent the negative calcium balance and decalcification of the bones. Surgical removal of the hypertrophied glands or adenoma soon causes considerable improvement in the patients' condition. Removal of a functional parathyroid tumor

often causes hypocalcemia and tetany. This has been attributed to inhibition of the remaining parathyroid tissue or to the rapid removal of calcium from the circulation and its storage in the bones. Blood calcium should be carefully watched after the operation and hypocalcemia should be treated immediately in order to avoid tetany.

The mechanism of hyperparathyroidism. According to Albright, the initial disturbance in hypoparathyroidism is an excessive renal excretion of phosphorus, which would provoke mobilization of calcium and phosphorus in the bone, leading to decalcification. Parathyroid hormone, however, acts directly on the bone, as is shown by the fact that it provokes hypercalcemia and decalcification of bone in nephrectomized animals.

Secondary hyperparathyroidism. Hypertrophy of the parathyroids is found in several conditions in which there is a decrease in body calcium: (*a*) in diets with low calcium content (rats and rabbits); (*b*) in diets deficient in vitamin D (chicks); (*c*) after renal operations (rats); (*d*) in many patients with rickets and osteomalacia. In all these cases hypertrophy of the parathyroids helps to keep a normal blood-calcium level and lowers the blood phosphate. Evidently there is a close relationship between the parathyroids, the kidneys, and the bones. Hyperparathyroidism causes lesions in the bones and kidneys, and reciprocally renal lesions may cause secondary hyperparathyroidism. Albright and Ellsworth have described cases in which renal insufficiency causes first a retention of phosphate; this lowers the blood calcium, and the hypocalcemia provokes compensatory hypertrophy of the parathyroids.

BIBLIOGRAPHY

Albright, F., The Parathyroids, *J. A. M. A.*, **117**, 527, 1941.

Albright, F., and E. C Reifenstein, "The Parathyroi Glands and Metabolic Bone Disease," Williams & Wilkins, Baltimore, 1948.

Bensley, S. H., The Normal Mode of Secretion in the Parathyroid Gland of the Dog, *Anat. Rec.*, **98**, 361, 1947.

Collip, J. B., The Parathyroid Glands, *Medicine*, **5**, 1, 1926.

Greep, R. O., The Physiology and Chemistry of the Parathyroid Hormone, in "The Hormones," Vol. 1, Academic Press, New York, 1948, p. 255.

Lewis, J. T., "Las Glándulas Paratiroides," El Ateneo, Buenos Aires, 1932.

McLean, F. C., Activated Sterols in the Treatment of Parathyroid Insufficiency, *J. A. M. A.*, **117**, 609, 1941.

Morton, J. J., Las Paratiroides, in "Tratado de Endocrinología Clínica," Vol. 2, G. Kraft, Buenos Aires, 1950, p. 9.

Shelling, D. H., "The Parathyroids in Health and Disease," Mosby, St. Louis, 1935.

Snapper, I., "Medical Clinics on Bone Diseases," Interscience Publishers, Inc., New York, 1943.

Thomson, D. L., and J. B. Collip, *Physiol. Rev.*, **12**, 309, 1932.

CHAPTER 56

Pancreas. Thymus. Epiphysis

PANCREAS

The pancreas is an endocrine gland of great importance owing to its secretion of insulin. The physiologic significance of insulin has been studied together with carbohydrate metabolism in Chap. 41. The part it plays in fat metabolism has been studied in Chap. 42.

THYMUS

There is no final proof that the thymus is an endocrine gland; the work that has been done on this problem has so far given contradictory results.

The thymus has a typical growth cycle; it increases in size and weight up to eleven or twelve years of age, and at fifteen years begins to retrogress, the thymocytes and reticular cells being gradually replaced by fat. During pregnancy there is also a typical involution of the thymus, which recovers its size after delivery. The thymus suffers involution in many circumstances, such as inanition and infections, and by the action of x-rays or certain hormones. Autopsies performed on subjects who have died of an infectious or wasting disease, even if they are young, reveal an atrophic thymus. On the other hand, subjects who have died suddenly, *e.g.*, owing to an accident, have an apparently large thymus.

Action on growth. Extirpation of the thymus has often been performed in many species. Thymectomized birds have a tendency to lay eggs without a shell (Soli, Riddle). In the adult mammal, thymectomy produces no symptoms. In young mammals disturbances in growth have been observed after removal of the thymus. Retarded body growth and sexual maturation, decalcification of bones or rickets, mental deficiency, and cachexia have been reported in thymectomized animals. On the other hand carefully controlled experiments in rats (Chiodi) and in dogs (Park and McClure) have given negative results, *i.e.*, the animals developed without showing any abnormalities in general body growth or sexual development. In many of the experiments in which retardation of growth was observed. the diet was not adequately controlled, and the results may have been due to dietary deficiencies.

It has been said that if the thymus is removed at an early age in successive generations of white rats, after several generations there is a retardation in growth (Einhorn and Rowntree). This remarkable observation has not been confirmed by other workers (Chiodi *et al.*) who performed carefully controlled experiments. Thymus extracts that cause an acceleration in growth have been obtained (Asher). Rowntree and his associates[1] have reported that treatment with thymus extract for several successive generations of rats accelerated growth and development. The animals were born with a significantly higher weight than the controls, with well-developed teeth, and covered with hair, and the eyes opened a few hours after birth; sexual maturity was also accelerated. Later the rate of growth decreased, and the final size and aspect of the animals were normal. These results have not been confirmed by others (Chiodi *et al.*).

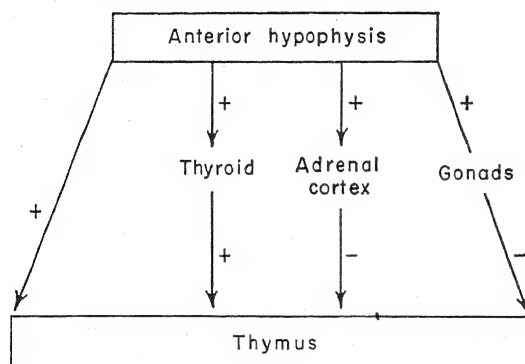
Action on metabolism. The thymus has been considered a factor in nucleoprotein and carbohydrate (Bomskow) metabolism, but this has not been confirmed.[2]

The Control of the Thymus by Endocrine Glands

The thymus is controlled by several endocrine glands. The gonads and the adrenal cortex inhibit it; the thyroid has a stimulating effect.

[1] Rowntree, L. G., G. A. Clark, A. M. Hanson, and A. Steinberg, *J. A. M. A.*, **103**, 1425, 1934.

[2] Castellanos, H., *Rev. Soc. argent. de biol.*, **23**, 51 and 56, 1947.

The anterior hypophysis has a direct stimulating effect on the thymus due to the growth hormone, and an indirect effect due to the secretion of thyrotrophin which stimulates thyroid secretion. The anterior hypophysis also has an indirect inhibitory effect on the thymus due to the secretion of adrenocorticotrophin, which stimulates the adrenal cortex, and of gonadotrophins, which stimulate the sexual glands.

Gonads. The growth curve of the thymus has the same shape in castrated animals as in normal ones, but at a higher level, *i.e.*, at the same age the thymus is larger in the castrates than in the controls (Fig. 259). Physiologic involution nevertheless takes place in the castrates in spite of the absence of the gonads; therefore it is not caused by the secretion of the sexual glands. The gonads exert a continuous inhibitory influence on the thymus, an influence that is removed by castration. An excess of male or female hormone causes atrophy of the thymus. Gonadotrophins also provoke involution of the thymus because they stimulate the secretion of the sexual hormones; they have no effect on the thymus of castrates.

Adrenals. Adrenalectomy is followed by hypertrophy of the thymus (Jaffe, Rapela, and others), which lasts only until the accessory adrenals have developed.[1] The condition of the thymus in cases of Addison's disease has been reported in only a few instances; sometimes early involution of the thymus was observed. It must be kept in mind that anorexia is an outstanding symptom of this disease, and that tuberculosis of the adrenals is its most frequent cause; therefore there are two nonadrenal factors (*i.e.*, undernourishment and infection) that cause atrophy of the thymus in Addison's disease. Hyperactivity of the adrenal cortex and injection of several of the corticoadrenal steroids provoke atrophy of the thymus. Adrenocorticotrophin has the same effect if the adrenals are intact, but not in adrenalectomized animals (Moon). The thymus atrophies in animals submitted to conditions of stress, such as cold, traumatisms, etc., but not if the animals have been adrenalectomized or hypophysectomized (Selye). In these cases the anterior hypophysis is stimulated and secretes an excess of adrenocorticotrophin, which causes hyperactivity of the adrenal cortex and thus involution of the thymus.

Thyroid. Thyroidectomy causes in most cases a rapid involution of the thymus (Marine, Chiodi, Pinto, and Reforzo). In cases of thyroid insufficiency in man, there are few reports on the state of the thymus. Involution of the thymus has been found (Wegelin), but it may have been due to nonthyroid causes, such as malnutrition. Thyroid treatment in adequate doses and in well-fed animals causes hypertrophy of the thymus which has been attributed to inhibition of adrenocorticotrophin secretion. Large doses of thyroid, especially if there is cachexia, cause atrophy of the thymus. Autopsies in human cases of hyperthyroidism often reveal the presence of a persistent thymus.

Hypophysis. Hypophysectomy is rapidly followed by involution of the thymus in the dog.[1] The hypophysectomized rat has a small thymus, even if it is forcibly fed by stomach tube to prevent the effects of underfeeding. Purified hypophyseal growth hormone causes hypertrophy of the thymus. A persistent thymus is often found at autopsy in cases of acromegaly.

Thymus and status lymphaticus. Paltauf described a pathologic entity, characterized by a large thymus and a highly developed lymphoid tissue, which was called "status lymphaticus." Most of the subjects in which this condition was reported had died suddenly. Hundreds of autopsies on persons who have died owing to an accident, or in war, have shown that when death occurs in healthy, well-nourished persons, the thymus and lymphoid tissues are well developed. The normal growth curve of the thymus has thus been obtained, and most, if not all, of the case histories of so-called "status lymphaticus"

[1] Houssay, B. A., E. B. del Castillo, and R. M. Pinto, *Rev. Soc. argent. de biol.*, **17**, 26, 1941; Rapela, C. E., Relación de la Glándula Suprarenal con el Timo, thesis for M.D., University of Buenos Aires, 1944.

[1] Houssay, B. A., and J. M Lascano-González, *Rev. Soc. argent. de biol.*, **10**, 241, 1934.

are merely descriptions of the normal condition of lymphoid tissue.

Thymus atrophy provoked by x-rays. The thymus is very sensitive to x-rays, which cause it to atrophy. X-rays act directly on the thymus and produce their effects not only in normal animals but also in hypophysectomized or adrenalectomized ones.

The thymus in myasthenia gravis. In more than half the cases of myasthenia gravis, hypertrophy or a tumor of the thymus has been reported. In some instances irradiation of the thymus with x-rays, or thymectomy, has improved the condition of the patients; in others it has had no effect.[1]

[1] Clagett, O., *Surg., Gynec. & Obst.*, **76**, 250, 1943; Viets, H. R., *J. A. M. A.*, **127**, 1089, 1945; Schwab, R. S., *et al.*, *Presse méd.*, **60**, 1501, 1952.

EPIPHYSIS

The epiphysis, or pineal body, develops from an area that gives rise to glandular tissue. Its histologic structure has been well studied by Del Rio-Hortega[1] and seems to be that of an organ with secretory activity. Experimental and clinical data on the epiphysis obtained up to now are contradictory, and without further evidence it cannot be considered as an endocrine gland. Precocious sexual development (early puberty) has been reported after extirpation of the epiphysis in cocks (Foà) and other animals (Sarteschi, Horrax, Martin, and Davis).[2]

[1] Del Rio-Hortega, P., *Trab. Labor. Invest. Biol. Madrid*, **21**, 95, 1923.

[2] Davis, L., and J. Martin, *Arch. Neurol. & Psychiat.*, **43**, 23, 1940.

Fig. 259. Effect of castration at 30 days of age on thymus weight. *a*, total weight of thymus of normal (lower curve) and castrated (upper curve) male rats; *b*, weight of thymus per 100 gm. of body weight in normal (lower curve) and castrated (upper curve) male rats; *c*, total weight of thymus of normal (lower curve) and castrated (upper curve) female rats; *d*, weight of thymus per 100 gm. of body weight of normal (lower curve) and castrated (upper curve) female rats. (*Chiodi, H., Rev. Soc. argent. de biol., vol. 14, p. 74, 1938.*)

Careful experiments by Dandy in dogs,[1] in the rat,[2] and in birds (Bardetscher) have been completely negative. In certain cases of tumor of the epiphysis in children, accelerated growth and precocious puberty have been reported (macrogenitosomia). In women with a pineal tumor, amenorrhea is usually observed. All these disturbances may be due to compression of the nerve centers in the neighborhood of the tumor and to a secondary effect on the hypophysis.

[1] DANDY, W. E., *J. Exper. Med.*, **22**, 132, 1915.
[2] DEL CASTILLO, E. B., *Rev. Soc. argent. de biol.*. **4**, 204, 1928.

BIBLIOGRAPHY

ANDERSEN, D. H., The Relationship between the Thymus and Reproduction, *Physiol. Rev.*, **12**, 1, 1932.

CHIODI, H., El Timo en Relación con el Crecimiento y la Función Sexual, thesis for M.D., University of Buenos Aires, 1938.

CHIODI, H., The Relationship between the Thymus and the Sexual Organs, *Endocrinology*, **26**, 107, 1940.

HAMMAR, J. A., "Die normal morphologische Thymusforschung in letzten Vierteljarhundert. Analyse and Synthese nebst einigen Worten zu der Funktionfrage," Barth, Leipzig, 1936.

RAPELA, C. E., Relación de la Glándula Suprarenal con el Timo, thesis for M.D., University of Buenos Aires, 1944.

SECTION SEVEN

Reproduction

CHAPTER 57

Sexual Characters and Their Regulation

REPRODUCTION TAKES PLACE in vertebrates when the ovum (female gamete, produced by the ovary) is fertilized by the spermatozoon (male gamete, produced by the testicle).

Sexual characters. There are many characters that differentiate the male from the female of the species. There are primary and secondary sexual characters. The primary sexual character is the organ that produces the gametes, *i.e.*, the testicle in the male and the ovary in the female (the gonads). Secondary sexual characters (the name was first used by Hunter in 1780) may be classified into morphologic, functional, and psychic. Morphologic sexual characters are divided into genital and extragenital. Genital secondary sexual characters (or accessory organs of sex) are the ducts that convey the gametes outside the body, together with the glands associated with these ducts. In the male they are the epididymis, vas deferens, urethra, and penis, with the prostate, seminal vesicles, and Cowper's glands. In the female they are the fallopian tubes, uterus, vagina, vulva, and clitoris, and the glands in the mucosae of the female genital tract. Extragenital secondary sexual characters are all those (morphologic, functional or psychic) that differentiate the sexes.

It may be useful to mention a few outstanding secondary sexual characters in species commonly employed in the experimental study of sexual physiology.

Amphibians. The male of the toad *Bufo arenarum* (Hensel) is smaller than the female, and its skin is of a more uniform color, the female being of two well-marked shades of green; the male has a callus on the thumb and croaks when it is pressed under the arms. The male of the frog *Leptodactylus ocellatus* (L.) Gir. has large forelegs and a marked clasping reflex throughout the year.

Birds. The male characters in the domestic cock are the comb and other cephalic excrescences, crowing, aggressive behavior, and searching for and treading the females. The female characters in the hen are the plumage, sexual and maternal instincts, and inhibition of the development of the spurs (Fig. 260). The plumage is not a male character in cocks; capons and castrated hens have the same plumage as cocks.[1]

Mammals. The male is usually larger and has stronger muscles than the female; in certain species the male has cephalic appendages, such as horns (stag), a mane (lion) or beard (goat), or more developed teeth (boar) or incisors (elephant). In females the mammary glands, the marsupial pouch, and the sexual cycle are prominent characters that distinguish them from the males of the species. There are also many functional and psychic differences between the sexes.

The main extragenital secondary sexual characters in the human species are the following:

Man	*Woman*
Larger corporal development	Smaller corporal development
Greater stature	Lower stature
Rudimentary mammary gland	Well-developed mammary gland
Narrow pelvis	Broad pelvis
Greatest transverse diameter at shoulders	Greatest transverse diameter at hips
Less subcutaneous fat	More subcutaneous fat

[1] Sebright bantam cocks have "henny" plumage which is conditioned by the testes; capons have "cocky" plumage.

Man	*Woman*
Thick, rough skin	Thin, soft skin
Beard, mustache, tendency to baldness	No beard or mustache, hair on head longer and without tendency to baldness
Hair on pubis extending toward the umbilicus and the anus (lozenge shape)	Hair on pubis limited by horizontal line (triangular shape)
Abundant body hair, especially on the chest and limbs	Body hair fine and scarce
Well-developed larynx and low-toned voice	Less developed larynx, high-toned voice
Higher basal metabolic rate	Lower basal metabolic rate
Higher erythrocyte count	Lower erythrocyte count
Smaller adrenals and thyroids	Larger adrenals and thyroids, which also respond more readily to stimulation
Male character ot hypophysis	Female character of hypophysis
Male psychic character and response, initiative, aggressiveness, abstract thinking, idealism, interest in social problems	Female psychic character and response, more highly developed emotivity and affections, maternal and family sentiments, practical outlook

It is absurd to discuss the superiority of one sex over the other. There are some characteristics common to both sexes and others peculiar to each sex, which complement those of the other.

Sexual characters and functions evolve in the course of life. They are incompletely developed in the child before puberty. They grow rapidly during puberty and reach a maximum which is maintained during the years of maturity. Sexual activity declines in old age. In women there is a rapid retrogression which takes place in the course of the fifth decade of lite in the majority of cases, the most notable sign being cessation of the menstrual cycle. This is known as the menopause or climacteric. In man sexual decline takes place more gradually.

Sex is determined by genetic factors in vertebrates and in the majority of invertebrates. Nevertheless the gonadal germ of vertebrates is potentially bisexual during the first stages of its development, and several external and internal factors may exert an influence that can cause the sexual germ to become differentiated in the direction of either sex. Later, sexual differentiation is controlled mainly by the sexual hormones secreted by the gonads (ovaries or testes), which provoke and maintain the development of the secondary sexual characters corresponding to each sex.

DETERMINATION OF SEX

Sex is an inherited character. The sex of the fertilized egg is determined by its genetic constitution. Sexual differentiation is the result of this genetic constitution and other factors (internal and external) which act on the sexual germ. In man there are two types of sperm and only one type of egg (male digamety or heterogamy); in birds there are two types of egg and only one type of sperm (female digamety or heterogamy). In most insects there is male digamety, but in Lepidoptera (moths and butterflies) the female is heterogamous. Fertilization by one type of sperm in the human species produces an egg that develops into a male, and fertilization by the other type of sperm results in an egg that develops into a female. Human oogonia have 48 chromosomes (24 pairs); 46 are somatic chromosomes, and the other pair is known as the X chromosomes. During the maturation division of the germ cells the number of chromosomes is reduced to one-half (meiosis), because one member of each pair of chromosomes goes into each one of the daughter cells; therefore each ovum has 23 somatic chromosomes and 1 X chromosome. Spermatogonia also have 48 chromosomes—23 pairs of somatic chromosomes, and a pair formed by the X and Y chromosomes. In the maturation division the number of chromosomes is reduced to one-half, and two types of sperm are formed (male digamety or heterogamy). Half the spermatozoa have (23 + X), and the other half (23 + Y), chromosomes. Eggs resulting from fertilization by the former develop into females; they have (46 + XX) chromosomes; eggs resulting from fertilization by the latter have (46 + XY) chromosomes and develop into males.

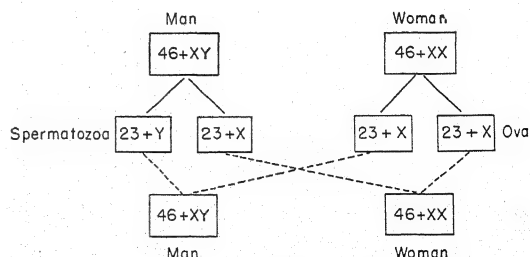
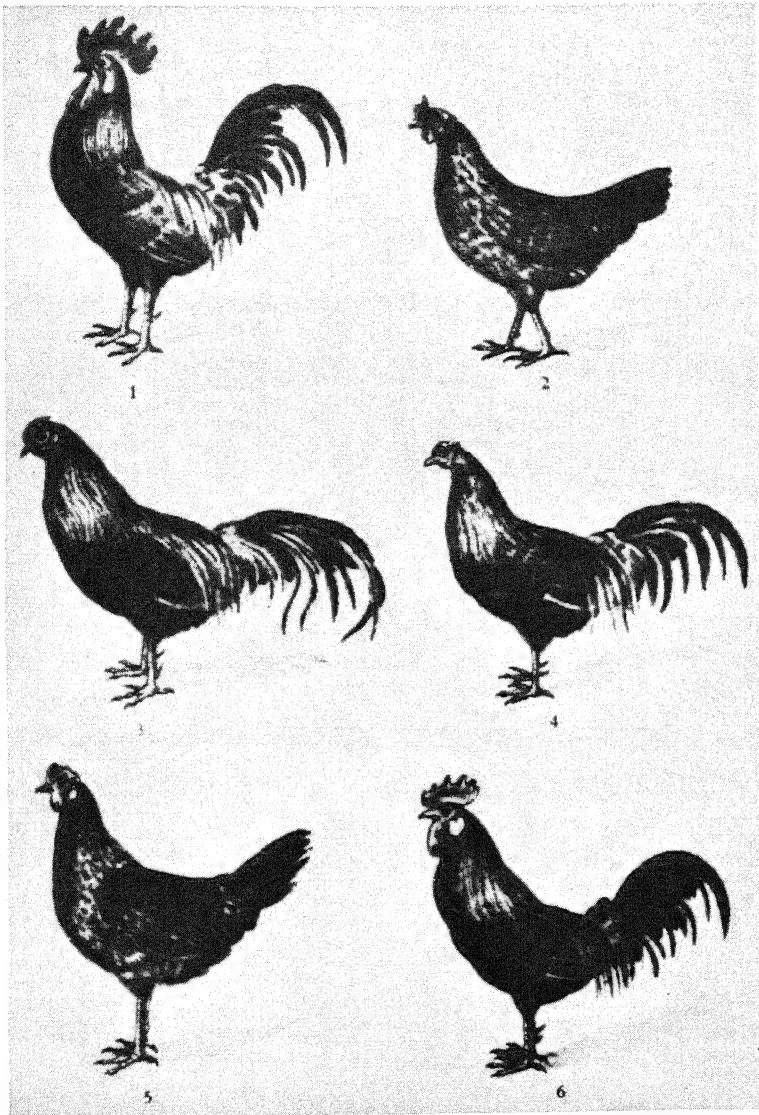



FIG. 260. Hormonal influence on sexual characters. The spurs have been slightly accentuated. 1, normal cock; 2. normal hen; 3, castrated cock; 4, castrated hen; 5, castrated cock with ovarian graft; 6, castrated hen with testicular graft. (*According to Zawadosky, from Martins, T., "Glandulas Sexuaes e Hypophyse Anterior," São Paulo, 1936.*)

If both types of sperm have the same fertilizing capacity and both types of egg survive in equal numbers, according to the laws of probability equal numbers of males and females will be conceived and born. In fact from 104 to 109 (average 106) boys are born for every 100 girls. In the course of life, however, the specific mortality for males is higher than for females (*i.e.*, men die at an earlier age than women). Therefore as age advances more women survive, and the sex ratio in the total population may show a predominance of women.

SEXUAL DIFFERENTIATION

Sex is determined genetically, but sexual differentiation takes place gradually in the course of development. The primitive sexual germ is potentially bisexual. In vertebrates usually the cortex of this germ has the capacity to develop into ovarian tissue, and the medulla into testic-

ular tissue. In the genetically female organism the cortex develops into ovaries and the medulla remains rudimentary or atrophies; in the genetically male organism the medulla develops into testes and the cortex atrophies.

Several internal and external factors condition the development of sexual differentiation. The most important internal factor is the sexual hormone, which stimulates the development of the gonad and sexual characters of the corresponding sex and inhibits those of the opposite sex.

Hormonal control of sexual differentiation. Three aspects of this control will be considered: (*a*) sexual differentiation of the gonads; (*b*) sexual differentiation of the genital tract, and (*c*) postnatal development of sexual characters.

Genetic and humoral factors control sexual differentiation of the gonads. According to Witschi the latter are (*a*) inductors of sexual differentiation, and (*b*) sexual hormones. The inductors are two: (*a*) medullarin, a protein which inhibits the development of the cortex (ovary) of the embryonic sexual germ, the medulla developing into a testicle; (*b*) cortecin, which inhibits the medulla, the cortex developing into an ovary. These substances have not been obtained in a pure state, and many workers (Dantchakoff, Wolff, etc.) consider that all the effects are produced by sexual hormones, which perhaps differ from those commonly used for experimental purposes (Jost). In certain species, such as cows and pigs, in cases of twin pregnancy with fetuses of different sexes, if there is placental fusion (*i.e.*, vascular communication between the placentas), the female fetus, which is called a freemartin, is masculinized and its ovaries are sterilized or converted into testes (Keller and Tandler; Lillie). This anomaly has been attributed to the secretion by the male fetus of male hormones, which pass through the communicating placentas into the female fetus. A genetic factor has also been postulated, especially as chorionic fusion may occur in other species in cases of twin pregnancy with fetuses of different sexes, without masculinization of the female (armadillo, man, etc.).

When two larvae of *Amblystoma* (axolotl) are joined in parabiosis, both always develop into individuals of the same sex (Burns), because the individual with the predominant sex modifies the gonad of the partner and causes it to differentiate into its own sex, whatever the original sex of the weaker partner may have been.

Sexual differentiation of the egg can be modified by the injection of sexual hormones into the allantoid (Dantchakoff). Male hormone converts the ovary of a female embryo into a testicle, and female hormone sometimes converts the testes of male embryos into ovariotestes. Injection of sexual hormones into the fetus or the pregnant female, in several mammalian species, causes the development of sexual characters corresponding to the hormone injected, and intersexual individuals or pseudohermaphrodites are developed.[2]

In the rabbit fetus, Jost[2] has shown that the internal secretion of the testes stabilizes or differentiates the wolffian ducts and structures developing from them and inhibits müllerian structures. The ovary, on the contrary, stimulates the development of the female genital tract, which, however, can grow even in the absence of the ovary. These facts were demonstrated by castration and grafting experiments in fetuses. Decapitation, which removes the hypophysis, did not prevent development, but the thyroid, adrenals, and testes were deficient in development and function.

Sexual reversal (*i.e.*, change of sexual characters) can be provoked in adult amphibians and birds by modifying the endocrine balance. These animals are potentially bisexual; one sex is evident, the other remains latent. The adult male toad has male characters, but it is a potential hermaphrodite, because attached to the well-developed and functioning testicle there is Bidder's organ, which is a rudimentary ovary. Removal of the testes causes the development of Bidder's organ (stimulated by the hypophysis)[3] into an adult ovary (Harms, Ponse). The opposite case, *i.e.*, conversion of a female into a male, is often seen in adult hens. If the left ovary, which is well developed, is extirpated, the remaining rudimentary right ovary is converted into a well-developed and functioning testicle in 10 per cent of the cases. The comb grows, the birds crow, female plumage is replaced by the neutral plumage of the cock at the first moulting, sexual and general behavior is that of cocks, and the birds tread hens and can fertilize them. Other cases of development of phenotypes opposite to the genotype of the animal will be considered when discussing pseudohermaphrodites and sexual hormones.

Modification of sexual characters by external factors. Tadpoles of certain amphibians, when kept

[1] Greene, R. R., *Cold Spring Harbor Symposia on Quantitative Biol.*, **5**, 105, 1937; Greene, R. R., M. W. Burrill, and A. C. Ivy, *Am. J. Anat.*, **65**, 415, 1938; **67**, 305, 1940; *Physiol. Zoöl.*, **15**, 1, 1942.

[2] Jost, A., *Biol. Rev.*, **23**, 201, 1948.

[3] Houssay, B. A., and J. M. Lascano-González. *Rev. Soc. argent. de biol.*, **7**, 248, 1931.

at 25 to 30°C., develop into males in a high proportion; at 10°C. females are predominant (Witschi). External temperature modifies the functions of the ovary and testes (see Chaps. 59 and 60).

HORMONAL DEVELOPMENT AND MAINTENANCE OF SEXUAL CHARACTERS

The development of sexual characters is conditioned by genetic and hormonal factors. Not only the sexual germ, but also the whole organism, is to a certain extent potentially bisexual and can respond to hormones of its own and of the opposite sex.

The development and control of sexual characters are studied by the following methods: (*a*) observation of the appearance of sexual differences in form, functions, and behavior of the male and female; (*b*) observation of the effects of prepuberal and postpuberal castration; (*c*) observation of reappearance of sexual characters in castrates by implantation or grafting of the gonads or by treatment with extracts or hormones of the sexual glands; (*d*) establishment of a state of sexual hyperactivity by overdosage with sexual hormones or extract, or observation of such states in pathologic cases.

The effects of castration. Castration produces certain effects which man has known for centuries and which have been used with the object of taming animals or modifying their flesh in order to increase its value as foodstuff. Berthold in 1849 observed that a testicular graft caused the development of male characters in capons, thus demonstrating the endocrine male function of the testes (Fig. 261).

In most invertebrates, sexual characters are apparently conditioned only by genetic factors, since castration does not modify them nor prevent their development. In vertebrates, on the contrary, sexual characters are fully developed and maintained only if the endocrine function of the ovaries or testes is normal. If the gonads fail to develop or castration is performed before puberty, sexual characters do not develop; if castration is performed in the adult when sexual differentiation is complete, some of the characters retrogress, others retain their male or female type. Pézard has distinguished prepuberal from intrapuberal and postpuberal castration; the effects are more marked the earlier castration is performed. Castration can be the result of surgical removal of the gonads, application of x-rays, injection of antagonistic hormones, nutritional deficiencies, etc.

a b c

FIG. 261. Heads of adult cocks. *a*, normal; *b*, castrated 155 days before, when 157 days old; *c*, castrated, with testicular graft made 62 days before. (*Courtesy of Dr. M. J. Copello.*)

In male toads and frogs, castration causes disappearance of the callosity on the thumb and of the clasping reflex. In the castrated cock the comb atrophies (Fig. 261), the birds no longer crow, and they lose their aggressive and specifically sexual behavior. The castrated hen changes its plumage for the neutral plumage of the cock and capon at the first moulting, and the spurs develop (Fig. 260).

Certain tissues require the continuous action of the sexual hormones to maintain their normal condition;

castration rapidly causes marked atrophy of these tissues. They are extremely sensitive to the action of sexual hormones in the normal animal, but more so in castrates, and they respond to even small doses of hormones by a rapid and marked development. For this reason this response is used for the study and assay of preparations of sexual hormones or of substances with an activity similar to that of these hormones. The capon's crest, or the seminal vesicles or prostate of rodents, are used for the assay of the male hormones, and for female hormones the vaginal epithelium or the uterus of rodents, or the vaginal epithelium or endometrium of women.

Restitution of sexual hormones. Disturbances caused by the lack of development of the sexual glands or by castration can be controlled by restitution of the sexual hormones. The principal methods used for this purpose are the following: (*a*) implantation of gonadal tissue; (*b*) injection of gonadal extract or hormones; (*c*) grafting of the gonads. Subcutaneous or intramuscular implantation of gonadal tissue produces only transitory effects, which are observed as long as the active principle in the implant is being reabsorbed. Injection of extracts of testicular or ovarian tissue was one of the first methods of restitution used; it led to the discovery of the active principles, but these extracts contain a large amount of inactive substances and others that may be harmful. Active principles that control the effects of castration have been obtained from gonadal tissue and from organic fluids. After their chemical structure was established, they were prepared by synthesis, and they are now in common use in therapeutics and for experimental purposes.

Ovarian or testicular grafts control the effects of castration after they have "taken" and as long as they function (Figs. 260 and 261). Ovarian grafts "take" with relative facility, but testicular tissue does not "take" easily, except in some species, *e.g.*, the rat, in which testicular grafting is frequently successful. Grafts "take" more frequently when the animal's own tissue is used (autografts). They "take" a little less frequently when tissue from animals of the same species is used (homologous grafts); in this case it is advisable to use, if possible, animals of the same strain. Grafts do not "take" permanently when the tissue of a different species is used (heterologous grafts).

Grafted tissues "take" more easily in castrated than in normal animals, probably because there is hypersecretion of hypophyseal gonadotrophins in the former. They also "take" more readily in adult animals than in those which have not reached the age of puberty; in this case also extragonadal factors (*e.g.*, hypophyseal gonadotrophins) stimulate the development of the grafted tissue. Gonadal grafts develop and maintain secondary sexual characters, but they do not function as actively as the normal gland. The ovaries of guinea pigs can be kept in the refrigerator for a time before being grafted (Lipschutz). Rat ovaries have been kept at low temperatures in saline with 15 per cent glycerol and then grafted successfully (Parkes).

Often the grafted tissue at first shows signs of hyperactivity and later of hypofunction before it is reabsorbed.

Experiments have been made in which gonadal tissue is grafted into an animal of the opposite sex. These experiments have shown the soma can respond to both male and female hormones (Fig. 260). In other experiments a tissue that responds to stimulation by sexual hormones (*e.g.*, the comb of a cock, or the vagina or uterus of a rat, or mammary gland or skin) has been grafted into an animal of the same or the opposite sex. These tissues develop or atrophy according to the hormonal *milieu* into which they are transplanted.

THE EFFECTS OF SEXUAL HORMONES

Sexual glands exert their influence on the body by means of hormones which they secrete into the blood stream. Male and female hormones are steroids which are built around a perhydrocyclopentenophenatrene ring. Steroids are substances of great biologic importance; not only sexual hormones belong to this group, but animal steroids (cholesterol, etc.), plant steroids (phytosterols), bile acids, the D vitamins, corticoadrenal hormones, substances with a cardiotonic effect (digitalis), genins in toad venoms, odoriferous steroids, and cancerigenic substances also form part of this vast chemical family.

Hormones that stimulate the development and maintenance of feminine sexual characters are known as *estrogenic hormones* (from *oestrus*, sexual heat). Any substance that has a similar activity is called an *estrogen.* Hormones that stimulate the development and maintenance of

male sexual characters are known as *androgenic hormones* (Greek ἀνδρός, man or male), and *androgens* are substances that have the same effect.

Sexual hormones have stimulatory and inhibitory activities. Thus female hormones stimulate the development of female plumage and inhibit the development of the spurs in birds. They do not "create" organs or behavior, but merely modify the development of preexisting physical or psychic characteristics. Sexual hormones always produce more than one effect, but one of these may be more apparent than others. The action of sexual hormones is not evident until they have reached a certain concentration (threshold) in the body fluids. Sexual hormones are either excreted or destroyed fairly rapidly. They are not stored; therefore the signs of gonadal insufficiency are seen very soon after castration. The liver is the main organ in the process of inactivation of sexual hormones.

Sexual hormones act directly on the tissues and indirectly through the hypophysis.

Specificity. The sexual hormones have a specific effect. For instance, female hormones provoke the development of female characters in females, and some female characters in males; and male hormones provoke the development of male characters in males and to a certain extent in females (Fig. 260). These facts have been clearly demonstrated by experiments in which animals of both sexes have been castrated and have been treated by grafting gonadal tissue of the same or the opposite sex, or injecting sexual hormones of the same or the opposite sex. Specificity of effect (male or female) is a well-established fact, but sexual hormones are not zoologically specific, because they produce their effects in more than one zoological species.

Gonadal or hormonal bisexuality. Specificity of action of the gonads as stated above is a well-established fact, but the gonads can produce certain amounts of the hormone of the opposite sex. Moreover sexual hormones can have effects on the secondary sexual organs corresponding to the opposite sex. These are known as the heterosexual effects of the gonads or sexual hormones, which therefore, to a certain extent, have ambisexual or bisexual effects.

Among these heterosexual effects of the gonads, the following may be mentioned:

1. Both male and female sexual hormones are found in the urine of men and women; both hormones diminish considerably after castration. The urine of stallions is one of the richest sources of estrogens, which are also found in the equine testicle; castration diminishes the urinary excretion of estrogen in these animals. It has been supposed that sexual hormonal activity persisting in the urine after castration is due in part to the excretion of adrenal "corticoids."
2. The normal ovary produces very small amounts of androgens, but ovarian grafts made in the ear (Hill, 1937) or the tail (Hernandez, 1943) of the mouse produce them in much larger quantities.
3. Certain ovarian tumors called "arrhenoblastomas" produce substances that cause virilization (Fig. 263).
4. Ovarian grafts and estrogens attenuate or delay atrophy of the seminal vesicles in castrated males. Progesterone has a slight androgenic effect.
5. Testosterone has a slight action on the endometrium, vagina, and mammary gland.
6. High doses of estrogenic hormones cause estrus in the adult castrated male dog and rat. Androgenic hormones cause estrus in females when given in large doses.[1]

Antagonism between male and female gonads. There is an apparent antagonism between the gonads of the sexes due to the hypophysogonadal relationship,[2] but there is also a direct antagonism between the effects of the male and female hormones on the secondary sexual organs.

This antagonism is evidenced by the following facts:

1. Prolonged treatment with high doses of estrogens or androgens inhibits the gonadotrophic function of the hypophysis and causes ovarian or testicular atrophy. Estrogens are more efficacious in this respect than androgens.
2. The normal endocrine secretion of the gonads has an inhibiting effect on the secretion of gonadotrophins.
3. Castration is followed by an increase in gonadotrophin secretion, which favors the development of grafted gonadal tissue; on the other hand this tissue does not "take" easily in normal animals.
4. If an ovarian graft "takes" in an intact male guinea

[1] Beach, F. A., *Physiol. Rev.*, **27**, 240, 1947; "Hormones and Behaviour," Hoeber, Inc., New York, 1948.

[2] Moore, C. R., and D. Price, *Am. J. Anat.*, **50**, 13, 1932.

pig, the mammary gland develops and secretes milk, while the testes continue to function (Sand).

Simultaneous or successive injections of male and female hormones are followed by the effects of both hormones. In some cases there is summation of the effects of the hormones of both sexes; thus injections of estrogens and testosterone in adequate proportions cause a much greater development of the seminal vesicles than can be obtained by injecting either of the hormones singly.

Antagonism between the hormones of opposite sexes can nevertheless be demonstrated in many cases. Thus testosterone diminishes or prevents several of the effects of estrogens, *e.g.*, metaplasia (cornification) of the prostatic utricle, testicular atrophy, increase in weight of the anterior hypophysis, cornification of the vaginal epithelium, etc. Reciprocally, estrone applied locally diminishes the effect of testosterone on the cock's comb. Estrogens and progesterone (the hormone of the corpus luteum) in certain proportions suppress each other's effects, but an adequate proportion of estrogen strengthens considerably the effects of progesterone on the endometrium or the mammary gland.

Effects of an excess of sexual hormones. Large quantities of hormones, secreted or injected, cause an abnormal response in the sexual organs. This is observed in some cases of ovarian or testicular tumor with hypersecretion of sexual hormones, or when the gonads are stimulated by an excess of gonadotrophin secreted by the anterior hypophysis or experimentally injected. Thus in castrates the anterior hypophysis secretes larger amounts of gonadotrophins than in normal animals, and ovarian or testicular tissue grafted into a castrate shows signs of hyperactivity for a certain length of time. For example, an ovarian graft in a castrated guinea pig secretes continuously an excess of estrogen, which causes a remarkable development of the mammary gland similar to that occurring in pregnancy, and may provoke the secretion of milk in amounts sufficient to nurse a suckling guinea pig (Steinach, Lipschütz, Giusti, etc.).

REGULATION OF GONADAL FUNCTIONS

The endocrine function of the gonads is accurately controlled, as is shown by the maintenance of the secondary sexual characters at a constant level of development and the regularity of the sexual cycles. In mammals the endocrine and reproductive functions of the sexual glands are regulated by extragonadal factors, of which the hypophysis is the principal. In normal conditions the gonads do not function at a maximum, a fact that is demonstrated by the considerable increase in sexual function following the administration of hypophyseal gonadotrophins. Constancy of gonadal activity is well demonstrated in experiments in which one ovary is removed; in the remaining ovary the number of follicles ripening at each cycle is the same as reached maturity in both ovaries before unilateral ovariectomy (Lipschütz's law of follicular constancy),[1] and the females bear the same number of offspring at each pregnancy. On the other hand, if one uterine horn is extirpated, they bear only half the normal number of offspring. A single ovary maintains the development of secondary sexual characters and the regularity of the cycles in the same way as the two ovaries of an intact female.

The onset and development of puberty are conditioned by extragonadal factors. An infantile ovary grafted into an adult organism is rapidly transformed into an adult ovary, but if it is grafted into a prepuberal castrate it will remain in the infantile condition.

The secretion of the sexual glands tends to reach a certain level and then to remain stable at that level. If subtotal gonadectomy is performed or testicular tissue is grafted into a castrate, the gonadal tissue either tends to atrophy completely, so that the animal has the characteristics of a castrate, or it grows sufficiently to maintain a full development of the male characters. There is an "all-or-nothing" effect; there are no stable states intermedaite between the castrate and the full male, as Pézard demonstrated in barnyard cocks. Nevertheless during the process of development of the gonads (or of a graft) certain secondary sexual characters appear before others; the former have a lower "threshold," *i.e.*, respond to smaller amounts of male hormone than the latter. Thus in cocks first the comb begins to develop, then the birds crow, and finally the sexual instinct appears (Pézard).

In certain cases the equilibrium between the hypophysis and the gonads is disturbed, and a stable state of hyperfunction of the ovary or testes is gradually established. Thus a condition

[1] LIPSCHÜTZ, A., *Bol. Soc. biol., Concepción*, 2, 3, 1928.

of permanent hyperfunction of the ovary can be provoked experimentally by several methods (fragmentation, grafting, ligature), in which there are continuous or prolonged and frequently recurring estrus, considerable uterine hypertrophy, etc. This state of hyperfunction is maintained by an increase in the secretion of hypophyseal gonadotrophins.

FIG. 262. Hermaphroditism. *a*, true lateral hermaphroditism, an ovary (*Ov*) with a fallopian tube (*F*) on one side, and a testis (*Ts*) with epididymis (*Ep*) and vas deferens (*V.D.*) on the other; well-developed uterine horns (*U.H.*) on both sides; *b*, unilateral hermaphroditism, an ovary on one side and an ovariotestis with vas deferens and fallopian tube on the other. (*Crew, "The Genetics of Sexuality in Animals," Cambridge, New York, 1927.*)

DISTURBANCES IN SEXUAL DIFFERENTIATION

Sex is determined by the genetic constitution of the individual at the moment of fertilization. In vertebrates the soma is, nevertheless, to a certain extent capable of developing both male and female characters if the appropriate factors come into play in the course of prenatal or postnatal development. Three types of cases must be considered:

1. Individuals in which the primary sex characters of both sexes, *i.e.*, ovarian and testicular tissues, are present and active at the same time. These are known as true hermaphrodites.[1]
2. Individuals that have either testes or ovaries, *i.e.*, are either male or female, but present secondary sexual characters corresponding to the opposite sex. These are known as pseudohermaphrodites, or intersexual individuals.[2]
3. Subjects in which there is a change from one sex to another, *i.e.*, sex reversal.

The majority of cases of disturbance in sexual differentiation are due to anomalies in development. Most of them have been reproduced experimentally by castration and injection of hormones or grafts of the gland. Thus an ovary grafted into a cock, or treatment with estrogens, causes the animal to develop female plumage (Fig. 260). If an ovary is grafted into an adult male guinea pig the animal has the aspect and behavior of a male and can fertilize females, but the mammary glands undergo hypertrophy similar to that observed in pregnant females; they secrete milk, and the males can suckle young guinea pigs (Sand). Castration, homosexual and heterosexual gonadal grafts, and implantation of sexual hormones have been successfully performed in fetuses.

True or biglandular hermaphroditism. The simultaneous presence of functioning testes and ovaries is the normal condition in certain animals, *e.g.*, some of the worms, but in mammals it is seldom observed, except in pigs (Fig. 262). Amphibians and hens can be considered as potential hermaphrodites, because the rudimentary ovary (Bidder's organ in the toad) or testes can develop and function in certain circumstances, as was explained in the paragraph on sexual differentiation. In man, according to Young,[1] only 20 out of all the cases published can be considered as adequately proved cases of true hermaphroditism, in which an ovary was found on one side and a testis or

[1] From the Greek myth of Hermaphroditus, the son of Hermes and Aphrodite, who when bathing became joined in one body with the fountain nymph Salmacis.

[2] "Intersexuality" is a term that has often been given different meanings. It is applied most frequently to the appearance of characters corresponding to one sex in individuals of the opposite sex. A female intersex shows secondary sexual characters intermediate between those of male and female type.

[1] YOUNG, H. H., "Genital Anomalies, Hermaphroditism and Related Adrenal Diseases," Williams & Wilkins, Baltimore, 1937.

an ovariotestis on the other. Cases published after the appearance of Young's work bring the total of well-authenticated hermaphrodites up to 38 (Selye). The testicle on one side is usually joined to a vas deferens, a prostate, and a rudimentary penis. The ovary on the other side is connected with a fallopian tube and uterus. There is definite proof that the testis in these subjects produces male and female hormones. In certain hermaphrodites (pigs, goats, and a few human cases) a hereditary factor seems to play a part in the causation of the anomaly.

According to Young, hermaphrodites can be classified into (*a*) true lateral hermaphrodites, with an ovary on one side and a testis on the other (Fig. 262*a*); (*b*) unilateral hermaphrodites, with an ovary on one side and an ovariotestis on the other (Fig. 262*b*); (*c*) true bilateral hermaphrodites with an ovary and testis on each side.

Pseudohermaphroditism. Certain individuals are either male or female, because they have either testes or ovaries, but they have secondary sexual characters of both sexes. This pseudohermaphroditism may be due to causes that act before birth (congenital pseudohermaphroditism) or after birth, even in adult life (acquired pseudohermaphroditism). If ovarian tissue is present, the individual is a female pseudohermaphrodite; if testicular tissue is present, the individual is a male pseudohermaphrodite.

The congenital female pseudohermaphrodite has an ovary but the organs that develop from the müllerian duct (fallopian tubes, uterus) are usually rudimentary. There is no testicular tissue, but the clitoris is hypertrophied and in some cases has the size of a penis. The labia take on the aspect of a scrotum, and anomalies in the vaginal canal are frequently observed. The organs arising in the wolffian duct (vas deferens, etc.), do not develop. The mammary glands are frequently underdeveloped, there is hirsutism in most cases, and usually these individuals do not menstruate. These anomalies are probably due to the production of androgens by the ovary or the adrenal cortex in the course of fetal development (Jost).

The congenital male pseudohermaphrodite has a testicle (sometimes ectopic) with a vas deferens and prostate. The external genitalia have the appearance of the vulva. In some cases there is a uterus with fallopian tubes and a vagina, and other female characteristics, such as the shape of the body, mammary development, treble voice, and feminine behavior. The testicle apparently secretes female hormone; in some reported cases, the female characters have been observed to retrogress after the testicle has been extirpated. Jost suggests that these disturbances are probably due to prenatal deficiency in testicular hormone, and the secretion of feminizing factors.

Male acquired pseudohermaphroditism may be due to (*a*) feminizing tumors, such as chorioepitheliomas of the testes and exceptional cases of feminizing corticoadrenal tumors; (*b*) prolonged administration of high doses of estrogens. The external genitalia retrogress. There is testicular atrophy and abnormal development of the mammary gland (gynecomastia), occasionally with a pigmented areola.

Female acquired pseudohermaphroditism is usually due to hyperplasia or tumors of the adrenal cortex which

a b c

FIG. 263. Masculinization due to arrhenoblastoma in a twenty-nine-year-old woman. *a* and *b*, before operation, hirsutism and atrophy of the breasts; *c*, 6 months after the tumor was removed, abnormal hair has disappeared and the breasts have developed. Later the patient gave birth to two normal children. (*Althabe, A., D. Colillas, and G. Di Paola, Bol. Soc. de obst. y ginec. de Buenos Aires, vol. 18, p. 739, 1939.*)

secrete androgenic steroids, or to arrhenoblastomas. The latter are ovarian tumors, probably developed from the medulla of the ovary. They secrete a substance that exerts a powerful masculinizing influence resulting in hypertrichosis (abnormal development of the beard and body hair), masculine shape of the body, bass voice, etc. (Fig. 263). Male characters usually retrogress after the tumor has been removed.

Masculinization has been reported in a few cases of tumors developed from the ovarian theca cells (Fraenkel).

Hyperplasia and certain tumors of the adrenal cortex can produce pseudohermaphroditism in women, causing virilization with hypertrichosis (adrenogenital syndrome; see Fig. 252 and discussion in Chap. 54). Extirpation of the tumor is usually followed by retrogression of the male characters.

Sex reversal. Change of sex in the adult has been observed in certain fishes which at some times produce spermatozoa and at others produce ova. In some old hens the ovary atrophies, the rudimentary wolffian structures develop, and the hen is converted into a fertile cock (Crewe).

Gynandromorphism. In birds and insects individuals have been found that are a mosaic of male and female characters, *i.e.*, certain parts of the body are male, while others clearly distinguishable from the former are female; *e.g.*, one half of the body has a testicle and male characters, the other half an ovary and female characters. No cases of this anomaly have been observed in mammals.

BIBLIOGRAPHY

ALLEN, E., "Sex and Internal Secretions," 2d ed., Baillière, London, 1939.

CENTRE NATIONAL POUR LES RECHÉRCHES SCIENTIFIQUES, COLLOQUE INTERNATIONAL, "La Différentiation sexuelle chez les vertébrés," Paris, 1951.

JOST, A., Le Controle hormonal de la différentiation du sexe, *Biol. Rev.*, **23**, 201, 1948.

JOST, A., Sur le controle hormonal de la différentiation sexuelle du lapin, *Arch. d'anat. microsc. et de morphol. exper.*, **39**, 577, 1950.

JOST, A., Problems of Fetal Endocrinology: The Gonadal and Hypophyseal Hormones, *Recent Progress in Hormone Res.*, **8**, 379, 1953.

MOORE, C. R., "Embryonic Sex Hormones and Sexual Differentiation," Charles C Thomas, Springfield, Ill., 1947.

MOORE, C. R., The Role of the Fetal Endocrine Glands in Development, *J. Clin. Endocrinol.*, **10**, 942, 1950.

PÉZARD, A., and F. CARIDROIT, Caractères sexuels secondaires, in G. H. Roger and L. Binet, "Traité de physiologie normale et pathologique," Vol. 11, 2d ed., Masson et Cie, Paris, 1934, p. 163.

PONSE, K., "La Différentiation du sexe et l'intersexualité chez les vertébrés," F. Ronge et Cie, Lausanne, 1949.

TURNER, C. D., "General Endocrinology," Saunders, Philadelphia, 1949.

WITSCHI, E., Gonad Development and Function, *Recent Progress in Hormone Res.*, **6**, 1, 1951.

YOUNG, H. H., "Genital Anomalies, Hermaphroditism and Related Adrenal Diseases," Williams & Wilkins, Baltimore, 1937

CHAPTER 58

Endocrine Regulation of Sexual Functions

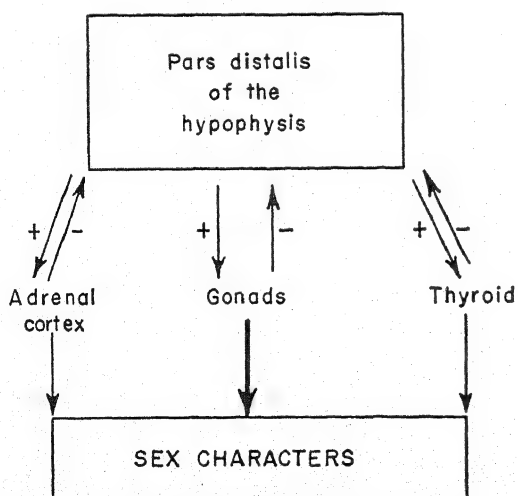
The endocrine sexual system. The sex glands have a double function: (*a*) *gametogenic, i.e.*, they produce and release the gametes of the corresponding sex, *i.e.*, sperm or eggs; (*b*) *hormonal, i.e.*, they secrete hormones that contribute to develop and maintain the secondary sexual characters, *i.e.*, the morphologic (genital and extragenital), functional, and psychic characteristics that differentiate one sex from the other. Among the psychic characteristics one of the most important is the libido, or sexual impulse toward the opposite sex.

The gametogenic and hormonal functions of the sex glands are governed by the pars distalis of the hypophysis, which secretes the gonadotrophins needed for the development and maintenance of the functions of the sex glands (see diagram). The hypophysis is therefore an organ of primary importance for reproduction.

Pars distalis of the hypophysis

+ − + − + −

Adrenal cortex Gonads Thyroid

SEX CHARACTERS

Reciprocally the hormones secreted by the gonads regulate the secretion of gonadotrophins, in some cases stimulating this secretion, but usually inhibiting it.

The adrenals have direct and indirect effects on the sex organs. The direct effects are due to the virilizing or feminizing activity of the corticoadrenal hormones. The indirect effects are exerted through the anterior hypophysis.

The thyroid also has direct and indirect effects on the sex organs. The indirect are due to the effect produced by thyroid secretion on gonadotrophin secretion.

Gonadotrophin secretion is regulated by the hormones of the gonads, adrenals, and thyroid, and in some cases by the activity of the nervous system.

Certain organs which exist only for a short period, such as the corpus luteum and the placenta, also secrete hormones of great importance in the process of reproduction.

In summary: (*a*) the pars distalis of the hypophysis regulates the development and functions of the sex glands; (*b*) five endocrine glands, *i.e.*, the testicle, ovary, placenta, and in a lesser degree the adrenal cortex and thyroid, secrete hormones acting directly on the secondary sex characters; (*c*) these five glands regulate the activity of the pars distalis. The gonads control not only the hypophyseal secretion of gonadotrophins, but also the secretion of adrenotrophin, thyrotrophin, and mammotrophin; thus indirectly they control the development and functions of these glands.

Several nutritive factors (*e.g.*, dietary deficiencies, avitaminoses, etc.), metabolic disturbances (*e.g.*, diabetes), and infectious diseases

can disturb the endocrine regulation of sexual functions by their effects on the hypophysis or by a direct action on the sexual organs.

HYPOPHYSIS AND SEXUAL FUNCTIONS

The pars distalis of the hypophysis is the fundamental factor in the reproductive functions of both sexes. It has a direct effect on the sex glands, which consists in (*a*) stimulation of the maturation of the gametes (spermatozoa and ova) and their release (formation of sperm and ovulation); (*b*) development of the gonads and maintenance of their structure and internal secretion. It has an indirect effect on secondary sex characters, mainly through the control of the secretion of sex hormones by the ovaries and testes, and to a slight degree through the control of corticoadrenal secretion of sex hormones. The latter are of little importance in normal individuals, but in certain abnormal conditions (adrenogenital syndrome) they exert a considerable influence on sex characters.

Hypophyseal gonadotrophins do not act directly on the external genitalia, vagina, uterus, and tubes, or on the penis, prostate, and seminal vesicles. They exert their influence on these organs exclusively by controlling the secretion of sex hormones by the gonads; therefore hypophyseal gonadotrophins have no effect in castrates.[1]

In mammals the anterior hypophysis is a necessary factor for the continuance of pregnancy; it exerts its influence on the processes of ovulation, implantation of the embryo, and maintenance of the embryo in the uterus. It is also a necessary factor in lactation; it contributes directly and indirectly to the development of the mammary glands and provokes and maintains the secretion of milk.

a *b* *c*

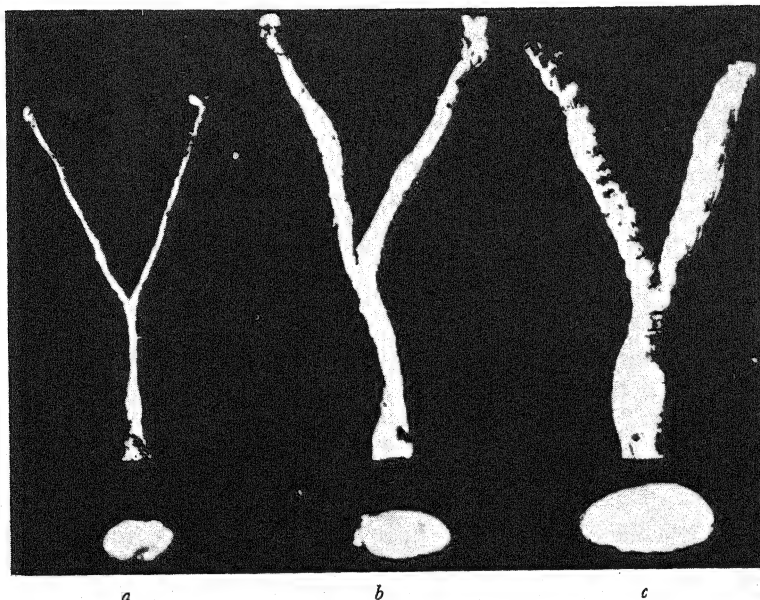



FIG. 264. Action of the hypophysis on the uterus and the ovary. *a*, hypophysectomized bitch weighing 9 kg.; the ovaries weighed 243 mg.; *b*, normal bitch, weighing 5.6 kg.; the ovaries weighed 395 mg.; *c*, bitch weighing 4.7 kg. after injection of anterior hypophyseal lobe extract; the ovaries weighed 1,192 mg. (The ovaries and the uterus have not been photographed with the same enlargement.)

The reciprocal influence of the hypophysis and the gonads has been demonstrated by the results of hypophysectomy, castration, and the injection of hypophyseal and sex hormones.

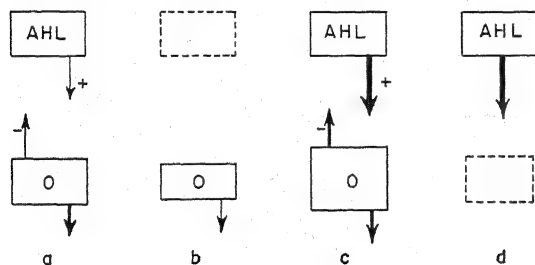
The effects of hypophysectomy on the sex glands. Hypophysectomy is rapidly followed by marked atrophy of the sex glands and consequently of the secondary sex characters, which

[1] There are a few exceptions to this general rule. Thus the anterior hypophysis stimulates the secretion of the oviduct in the toad *Bufo arenarum* (Hensel), even in castrated animals (Inés L. C. de Allende, 1938). Prolactin, secreted by the anterohypophysis, has a direct effect on the mammary gland.

are controlled by the internal secretion of the gonads[1] (Fig. 264). The primary follicles of the ovary start to develop, but when they reach a certain size they retrogress, and the ovary gradually loses its power to respond to hypophyseal hormones. The interstitial cells also retrogress in those species which possess this type of cell. The existing corpora lutea undergo involution, and new ones are not formed.[2]

Hypophysectomy prevents nidation of the egg and provokes abortion in many species (see "The hypophysis during pregnancy," page 661). Hypophysectomized animals do not secrete milk, and its secretion is suppressed by hypophysectomy in lactating females.

In males, hypophysectomy causes marked atrophy of the gametogenic and interstitial tissues of the testes. Spermatogonia continue to divide for some time after hypophysectomy, but spermatozoa are not formed; eventually all spermatogenetic activity ceases.

The effects of castration on the hypophysis. Castration suppresses the inhibitory effect of the sexual hormones on the anterior hypophysis; therefore gonadotrophins contained in, and secreted by, the pars distalis increase gradually in the course of several months and reach a considerable amount (see diagram).

a, normal reciprocal relationship between the anterior lobe of the hypophysis (AHL) and the ovary (O); *b*, hypophysectomy, ovarian atrophy and insufficiency; *c*, excess gonadotrophin, ovarian hypertrophy and hyperfunction; *d*, castration, excess secretion of gonadotrophin.

The gland increases in weight, and cytologic changes, which vary in different species, are observed. In the castrated rat the most typical change is the formation of vacuoles in the basophil cells, which gradually merge into a single large vacuole surrounded by a thin layer of protoplasm. The whole cell, with a prominent nucleus, has the aspect of a seal ring.

[1] Crowe, Cushing, and Homans, 1910; Aschner, 1913; Ascoli and Legnani, 1912; Houssay, 1916; Smith, 1927 to 1930.

[2] In the hypophysectomized rat, involution of the corpora lutea proceeds at a slower rate than in normal rats, but this is not seen in other species.

The increase in gonadotrophins can be demonstrated in several ways. Thus injection of extract of or implantation of the anterior hypophysis of castrates causes much greater effects than the same amount of normal anterior hypophyseal tissue. The increase in gonadotrophins in the blood is very clearly demonstrated by experiments of parabiosis (Martins). A male or female castrated rat is joined to a normal rat of the opposite sex. The abnormally large amount of gonadotrophins secreted by the castrate passes into the normal partner and stimulates the development of its sexual glands. If the normal partner is a female, whether adult or prepuberal, the ovaries are considerably enlarged by the ripening of numerous follicles and the formation of corpora lutea (Fig. 265).[1] The ovaries thus stimulated secrete large quantities of estrogenic hormone which cause hypertrophy of the uterus and cornification of the vaginal epithelium (permanent or continuous estrus). If the normal partner is a male, the testes are stimulated and there is hypersecretion of male hormone which provokes hypertrophy of the male organs, especially of the prostate and seminal vesicles (Fig. 265). Hypersecretion of gonadotrophins by the castrate hypophysis can be prevented or inhibited by grafting gonadal tissue or by injecting sexual hormone. Parabiosis of normal females does not cause any disturbance in the estral cycle of the partners. Parabiosis of normal males does not modify the sexual organs of the partners.

The effects of sexual hormones on the hypophysis. Injection of male or female hormone prevents the appearance of the cytologic changes in the anterior hypophysis and hypersecretion of gonadotrophins caused by castration, and it suppresses these effects if they have already appeared.

Female hormones stimulate the secretion of the luteinizing hormone and luteotrophin (prolactin) of the pars distalis, and corpora lutea are formed.[2] Adrenocorticotrophin secretion is also stimulated. Prolonged treatment with large doses of estrogens inhibits the secretion of all the anterohypophyseal hormones. Inhibition of the hypophysis is also obtained by the injection of male hormone, but estrogens have a more

[1] Matsuyama, 1921; Martins, 1929; Kallas, 1929; Fels, 1929.

[2] This effect of estrogens is, of course, not observed in hypophysectomized animals.

constant and powerful inhibitory effect than androgens.

The anterior hypophysis of these animals (rats) increases considerably in size; the blood vessels are dilated and give the gland a cavernous aspect. Only chromophobe cells are found; all the chromophil cells disappear. All the functions of the anterior hypophysis (gonadotrophic, thyrotrophic, adrenotrophic, etc.) are diminished. Sometimes a complete syndrome of anterior hypophyseal insufficiency is observed, *i.e.*, growth ceases and the gonads, thyroid, and adrenals atrophy as in hypophysectomized animals.

FIG. 265. Hypophysogonadal equilibrium. Parabiosis in rats. There is hypersecretion of gonadotrophin in the castrates of both sexes. 1, normal females; no change in estrus cycle or ovaries; 2, female castrate with normal female partner; in the latter there is hypertrophy of the ovaries and continuous estrus; 3, male castrate with normal female partner; in the latter there is hypertrophy of the ovaries and continuous estrus; 4, normal males; no change in seminal vesicles and prostate; 5, male castrate with normal male partner; in the latter there is hypertrophy of the seminal vesicles and prostate; 6, female castrate with normal male; in the latter there is hypertrophy of the seminal vesicles and prostate.

The male hypophysis differs from the female in its weight, cytologic aspect, and gonadotrophic and adrenotrophic activity. Thus the female anterior hypophysis secretes more adrenotrophin than the male, because estrogens stimulate this secretion and androgens inhibit it. These sexual differences in the hypophysis are produced by sexual hormones, therefore they must be considered secondary sexual characters.

The female anterior hypophysis shows periodic changes in activity and in its cytologic aspect, which coincide with the sexual cycle. The male hypophysis is more stable and secretes continuously. If testicular tissue is grafted into a newborn female mouse or rat, the hypophysis takes on a male aspect, which persists after the graft is removed. The masculinized hypophysis secretes gonadotrophins continuously, and the animals show endometrial changes typical of an excess of estrogen (Pfeiffer).[1]

The action of the nervous system on the hypophyseal-gonadal complex. Certain lesions in the hypothalamus provoke atrophy of the gonads and consequently retrogression of the secondary sexual characters.[2] In some cases there is also considerable accumulation of fat and a typical adiposogenital syndrome (Fröhlich syndrome) is observed. In these animals

[1] Lipschütz has observed that an ovary grafted into a male castrate guinea pig provokes greater development of the mammary gland and more abundant milk secretion than an ovary grafted into a female castrate. This is due to the secretion of estrogen provoked by continuous secretion of the hypophysis in the male; in the female periodically corpora lutea are formed and estrogen secretion diminishes.

[2] Aschner, 1912; Camus and Roussy, 1913; Houssay and Hug, 1921; Bailey and Bremer, 1921; Smith, 1930.

either all gonadotrophin secretion, or only the luteinizing gonadotrophin secretion, is inhibited (Dey).

Intense electrical stimulation of the hypothalamus (or of all the head) provokes a release of gonadotrophins which causes ovulation and luteinization in the rabbit and rat. This effect is not observed in hypophysectomized animals.

Intense and continuous illumination provokes the development of the gonads and secondary sexual characters in ferrets and birds in winter, when normally the animals are in a condition of sexual rest. This effect of light is not observed if the animals have the head or eyes covered or if they have been hypophysectomized.

In some species (rabbit, cat, ferret) ovulation does not occur spontaneously but is provoked by copulation. The female rabbit ovulates 10 to 18 hr. after copulation, but she does not ovulate if the hypophysis is removed before or up to ½ hr. after copulation (Fee and Parkes) or if the pituitary stalk has been cut (Brooks). In the female toad *Bufo arenarum* (Hensel) ovulation is provoked by the sexual clasp of the male (amplexus), but not if the pars distalis has been extirpated (Houssay). Pigeons ovulate when in the presence of another pigeon; a visual reflex must play a part in this process, because sometimes the bird ovulates if she sees her own image in a mirror.

The hypothalamus seems to exert some control on gonadotrophin secretion. In certain species its action can be suppressed by drugs (pentobarbital, atropine, dibenamine).

Section of the pituitary stalk suppresses some of the reflex effects on the hypophysis, such as ovulation in rabbits and the inhibition of estral cycles by cold in rats. However, most of the sexual functions of the hypophysis are not disturbed by this operation and the sexual cycle, estrus, copulation, ovulation, pregnancy, and lactation take place normally if vascular connections between the median eminence of the hypothalamus and the pars distalis are re-established (Harris and Jacobsohn). In these cases gonadotrophins are released and act without the hypothalamus taking part in the different processes. Gonadotrophins do not act through the hypothalamus in the process of ovulation, as is demonstrated by the fact that the gonadotrophins provoke ovulation in female toads in which the hypothalamus has been destroyed (Houssay and Giusti).

The effect of anterior hypophysis. Implantation[1] or grafts of the pars distalis prevent atrophy of the ovary or testicle after hypophysectomy, or restore the normal condition of the gonads if they are already atrophied. The gametogenic tissue and the interstitial cells are stimulated, and retrogression of the secondary sexual characters is prevented, or their normal condition is restored if they have already retrogressed.

Implantation of anterior hypophyseal lobe in infantile rats or mice causes precocious sexual development (precocious puberty). In adults it provokes intense sexual stimulation, and in old animals sexual activity is restored. These effects are obtained by implantation of the anterior hypophysis of the same or the opposite sex, and of the same or another species; *i.e.*, there is no specificity in this action of the hormones produced by the pars distalis.

Implantation of the pars distalis provokes the following effects in females: (*a*) abnormally abundant ovulation (superovulation), with the production of an abnormally large number of offspring; (*b*) increase in ovarian weight due to ripening and luteinization of a large number of follicles; (*c*) sometimes, pseudopregnancy, caused by the ovarian hypersecretion and luteinization; (*d*) congestion and hypertrophy of the uterus, and continuous estrus.

In the male, anterior hypophyseal implantation has the following effects: (*a*) development of the testes in prepuberal animals, but no increase in weight of the testes in adults; (*b*) hypersecretion of male hormone, which causes hypertrophy of the prostate, seminal vesicles, and other secondary sexual organs.

All these facts demonstrate that hypophyseal gonadotrophin secretion is controlled by nervous impulses only to a very limited extent. The hypophyseal-gonadal relationship is the most important factor in the regulation of gonadotrophin secretion; the thyroid and the adrenal cortex also play a part, and nutritive factors (*i.e.*, undernourishment, avitaminosis E, etc.) exert a certain influence on the functioning of the anterior hypophysis.

GONADOTROPHINS

Gonadotrophins are substances that stimulate the sex glands. Gonadotrophins are produced by the hypophysis and by the chorion. Hy-

[1] Ascheim and Zondek, 1926; Smith, 1926; Smith and Engle, 1927.

pophyseal gonadotrophins have been obtained from (*a*) hypophyseal extracts; (*b*) urine of women in menopause or after castration; (*c*) blood and urine of normal men and women. Chorionic gonadotrophins, *i.e.*, those produced by the placenta, have been obtained from (*a*) urine and blood of pregnant women; (*b*) blood of pregnant mares.

Hypophyseal gonadotrophins. Three gonadotrophins have been extracted from the pars distalis of the hypophysis: (*a*) the follicle-stimulating hormone; (*b*) the interstitial-cell-stimulating or luteinizing hormone;[1] (*c*) luteotrophin or prolactin. These substances are glycoproteins; the luteinizing hormone and luteotrophin have been obtained in a pure and crystalline form. They produce the following effects in mammals:

The follicle-stimulating hormone (FSH, or follicular maturation factor, or gametogenic hormone, or Van Dyke's thylakentrin) has a gametogenic action, *i.e.*, it provokes the formation of ova and spermatozoa.[2] The activity of this hormone is clearly demonstrated by its effect on hypophysectomized rodents (rats or mice). In the male it maintains or restores the normal condition of the seminiferous epithelium and the formation of spermatozoa, but it has no effect on the interstitial cells and does not stimulate the secretion of male hormone; therefore it has no effect on the secondary sexual characters. In the female it stimulates growth of the follicles but not of the interstitial cells, and does not stimulate the secretion of estrogens. In the normal female it provokes ripening and dilatation of the follicles.

The interstitial-cell-stimulating or luteinizing hormone[3] stimulates the interstitial tissue (Leydig's cells) in the normal and hypophysectomized male. It also stimulates the secretion of male hormone, with considerable development of the secondary sexual organs (prostate, seminal vesicles, etc.). Testosterone thus secreted stimulates the germinal epithelium and spermatogenesis. In the female this hormone converts the ripe follicles into corpora lutea; it stimulates the development of theca cells and the secretion of estrogenic hormones, but it does not prevent atrophy of the corpora lutea, nor does it stimulate their internal secretion. In the hypophysectomized rat it provokes growth of the involuted interstitial cells.

Luteotrophin (prolactin) maintains the structure and functions of the corpora lutea, even after hypophysectomy, and stimulates their internal secretion in the rat.

The association of follicle-stimulating hormone and luteinizing hormone in adequate proportions usually reinforces their effects (hormonal synergy). On the other hand, if the proportions of the hormones are not adequate they may inhibit each other's effects (hormonal antagonism). Hormonal synergy can be well demonstrated in normal and hypophysectomized animals by the development and ripening of the follicles, ovulation, and luteinization, with a great increase in ovarian weight. The greatest effects are obtained by injecting first FSH and then LH. The secretion of estrogens is increased, but the internal secretion of the corpora lutea is not stimulated.

Hormonal synergy is also observed in the male; the testes of infantile animals increase in weight, although this does not occur in adults; spermatogenesis is stimulated in both young and adult animals, and secondary sexual organs are considerably hypertrophied.

The follicle-stimulating hormone never produces luteinization in hypophysectomized females; but in females with a normal hypophysis FSH produces luteinization because it increases the secretion of LH by the anterior hypophysis and hormonal synergy comes into action.

The urine of normal women sometimes contains gonadotrophins (0 to 25 rat units per day); there are larger amounts (5 to 360 rat units daily) in the urine of women at the menopause.[1] The hypophysis undergoes cytologic changes at the menopause, and FSH (but not LH) is found in the blood and urine. Urinary and blood gonadotrophins, found in normal adult women

[1] These two hormones were first separated by Fevold, Hisaw, and Leonard (1931) and obtained in pure form by Li (1949).

[2] Highly purified FSH has only a small effect on the germinal epithelium; apparently it must be slightly contaminated with luteinizing hormone to produce its maximum effect.

[3] LH (luteinizing hormone, Fevold); ICSH (interstitial-cell-stimulating hormone, Evans *et al.*); or metakentrin (Van Dyke).

[1] Determination of urinary gonadotrophins gives valuable information in many clinical conditions (Klinefelter, F. H., F. Albright, and J. C. Griswold, *J. Clin. Endocrinol.*, **3**, 529, 1943) They can be concentrated by precipitation with alcohol or by ultrafiltration (Gorbman, Jungck, etc.).

and at the menopause, are supposed to be produced by the hypophysis.

Placental or chorionic gonadotrophins. Gonadotrophins of placental origin have been found in (*a*) the urine and blood of pregnant women; (*b*) the serum of pregnant mares.

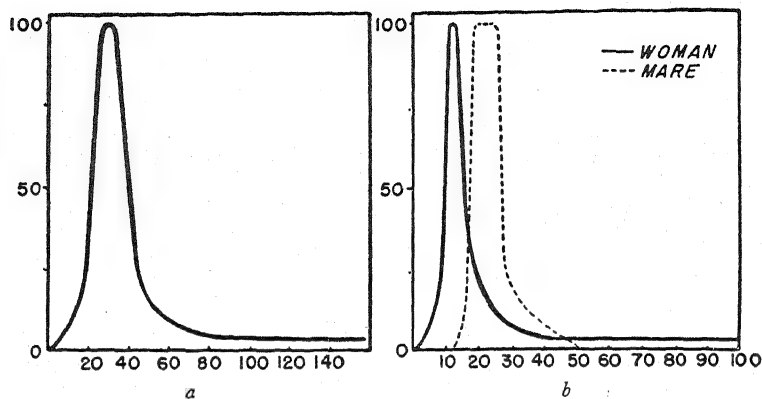



Fig. 266. Gonadotrophins in pregnant woman's urine. Comparison between the concentration of gonadotrophin in the urine of pregnant women (data from Evans *et al.*) and in the serum of pregnant mares (data from Cole *et al.*). *a*, duration of human pregnancy in days; *b*, percentage duration of pregnancy. Solid line, concentration of gonadotrophin in pregnant woman's urine; broken line, concentration of gonadotrophin in pregnant mare's serum. (*According to Evans, Kohls, and Wonder, 1937.*)

Pregnant women eliminate in the urine gonadotrophin produced by the placenta.[1] This gonadotrophin can be extracted from the placenta, and it is produced *in vitro* in cultures of placental tissue.[2] It is found in the urine of women who have aborted while the placenta is retained in the uterus, and it is eliminated in large quantities when there is a tumor of placental origin (hydatiform mole or chorioepithelioma) and in certain cases of tumor of the testicle.

Chorionic gonadotrophin has been purified and crystallized.[3] The crystalline substance has an activity equivalent to 6,000 or 8,000 units per milligram. It has a follicle-stimulating effect and a very strong luteinizing effect. It is found in blood and urine of pregnant women shortly after the nidation of the embryo. Its maximum concentration in the urine occurs around the thirtieth day of pregnancy (Evans), or 50 to 60 days after the last menstruation (Browne) (Fig. 266). Aschheim and Zondek utilized urinary elimination of gonadotrophin for the diagnosis of pregnancy (Aschheim-Zondek pregnancy test).

The serum of pregnant mares has a very active gonadotrophin (Cole and Hart, 1930), which is produced by special glandular structures of the placenta, the endometrial cups. It is a protein of great molecular weight; therefore it does not pass into the urine. It reaches its highest concentration between the fiftieth and eightieth day of pregnancy, then gradually diminishes, so that around the one hundred-eightieth day it has disappeared and is not found again for the rest of the 330 to 350 days of pregnancy. It produces effects even in hypophysectomized females, and in this respect differs from the gonadotrophin of pregnant women. It provokes follicular growth and luteinization.

The international unit of chorionic gonadotrophin is the specific activity of 0.1 mg. of the international standard gonadotrophin. The international unit of pregnant-mare gonadotrophin is the specific activity of 0.25 mg. of the international standard. The international unit of luteotrophin (prolactin) is the specific activity of 0.1 mg. of the international standard.

Tests for the diagnosis of pregnancy. Tests used in the diagnosis of pregnancy are based on the fact that the placenta produces chorionic gonadotrophin which is excreted in the urine, where it can be easily detected.

The Aschheim-Zondek test is performed as follows: Mice, 3 to 4 weeks old, weighing 6 to 8 gm., are treated

[1] This gonadotrophin has been found in the urine of pregnant orangutans, chimpanzees, and macaques, during certain stages of pregnancy. So far it has not been found in the urine of pregnant females of other species.

[2] Gey, Segar, and Helman, 1938; Stewart *et al.*, 1948.

[3] Claesson, L., B. Högberg, T. Rosenberg, and A. Westman, *Acta endocrinol.*, 1, 1, 1948.

during the course of 3 days with 6 injections of 0.3 to 0.4 cc. of the urine voided in the morning by the patient. It is useful to acidify the urine and wash it with ether. The animals are killed 100 hr. after, and the genital tract is examined. The following results may be observed in sections of the ovary (Fig. 267): (*a*) follicular dilatation; (*b*) hemorrhagic points (intrafollicular hemorrhage); (*c*) luteinization. Hemorrhage and luteinization indicate pregnancy; simple follicular dilatation does not. This test gives accurate results in 98.5 to 99 per cent of the cases.

Hypophysectomized rodents do not respond, or give only a rudimentary response (slight luteinization of the granulosa without ripening of the follicles), to the injection of urine from pregnant women. The results observed are usually due to the synergic effects of the chorionic gonadotrophins injected and gonadotrophins secreted by the animal's own hypophysis. There are several modifications of the Aschheim-Zondek test.

Friedman's test (Fig. 268) is the one of these modifications that is most commonly used, because it is completed in a shorter time and gives results that can be appreciated macroscopically. Mature female rabbits which have already had offspring or which weigh 1.8 to 2 kg. are used, after having been kept isolated for 4 weeks. If further precaution is desired the ovaries are examined before performing the test in order to be sure there are no hemorrhagic follicles. The urine to be tested (10 cc.) is injected slowly into the marginal veins of two rabbits. One animal is examined 24 hr. later and the other 48 hr. later. The test is positive (*i.e.*, there is pregnancy) if hemorrhagic follicles are found.

Other tests are based on the early appearance of congestion in the ovary, which can be detected by the naked eye under an intense illumination. These tests are not so reliable as the Aschheim-Zondek and Friedman tests; thus, 2 hr. after the injection the readings give 15 per cent of errors; 3.7 per cent when the

a *b*

c *d*

FIG. 267. Aschheim-Zondek pregnancy test. *a*, infantile ovary, immature follicles; *b*, response I, follicular dilatation, should not be considered positive; *c*, response II, hemorrhage in two very enlarged follicles, positive of pregnancy; *d*, response III, corpora lutea, positive of pregnancy.

observations are made 6 hr. after the injection, and 1 per cent when they are made 24 hr. after.[1]

Galli-Mainini's test[2] is performed as follows: A male toad (*Bufo arenarum*) weighing 100 gm. or more is injected subcutaneously with 10 cc. of the first urine voided in the morning. If the urine is that of a preg-

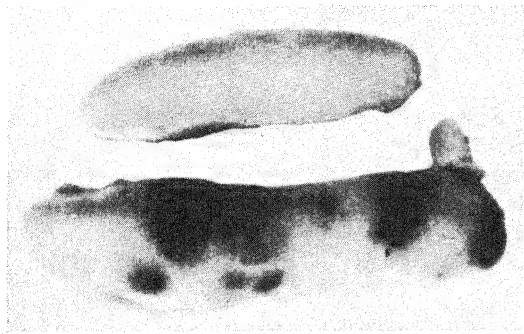



FIG. 268. Friedman's pregnancy test. Rabbits' ovaries. Above: mature follicles, no hemorrhage, negative. Below: ovulation and hemorrhage in follicles, positive of pregnancy.

nant woman, the Sertoli cells swell and release the spermatozoa attached to them. The spermatozoa are carried through the kidneys and ureters to the cloaca and are accumulated in the bladder. One to three hours after, the urine is collected from the cloaca by means of a pipet, and if the test is positive, numerous motile spermatozoa will be seen microscopically. Normally spermatozoa are never found in the cloaca except during amplexus (sexual clasping). The test is easily performed, the results are obtained within a very short time, and its cost is low. More accurate and rapid results can be obtained by injecting three animals and observing the cloacal contents of each one at alternate hours after the injection; sometimes there is a positive response at the first hour, but occasionally a response is not obtained for several hours. With urine from women in the first 5 months of pregnancy, 99.5 to 100 per cent positive tests have been obtained; with urine from women in the second half of pregnancy, only 92 per cent positive results have been observed. Other species of batrachians have also been employed successfully for the performance of this test.

The urine of pregnant women provokes ovulation in the South African toad *Xenopus laevis*. Zwarenstein and Duncan[3] have described a pregnancy test using this species. They have obtained 98.6 per cent accurate results in 1,407 tests.

Gonadotrophins in cases of malignant tumors.[1] Certain patients suffering from testicular tumors, especially chorioepitheliomas, eliminate large quantities of chorionic gonadotrophin (and sometimes of estrogens) in the urine. Other types of testicular tumor (seminomas) provoke the urinary excretion of gonadotrophin with the activity of the hypophyseal follicle-stimulating hormone. In some cases both types of gonadotrophin are excreted alternately or, more seldom, simultaneously.

The effects of gonadotrophins on women. The effects of hypophyseal gonadotrophins on women are not well known. Results obtained with gonadotrophins extracted from pregnant women's urine or pregnant mare's serum have not been encouraging; no effects, or no physiologic effects, were observed. It has not been possible to maintain the menstrual cycle permanently or to keep up a constant hormonal level. Prolonged treatment causes a decrease in sensitiveness to the preparation used, and sometimes the formation of antibodies occurs, especially when pregnant-mare-serum gonadotrophin is used.[2] Moreover, cases of allergic reaction have been reported.

Large doses of gonadotrophins provoke a transitory increase in the secretion of estrogens and the effects of hyperfunction of the corpus luteum.[3] By means of injections of large doses of pregnant-mare-serum gonadotrophin, followed by the injection of chorionic gonadotrophin, it has been possible to provoke menstruation in many cases of amenorrhea, ovulation has been obtained, and in a few cases pregnancy has resulted.

Large doses of crystallized chorionic gonadotrophin (36,000 units injected intravenously during the course of 3 days) provoke follicular stimulation without ovulation or luteinization. If this is combined with small doses (3,000 units) of pregnant-mare gonadotrophin, the follicles ripen and do not regress into atresia, but are converted into corpora lutea.

The effects of gonadotrophins on males. In some cases of ectopic testicle, injections of luteinizing or chorionic gonadotrophin cause the

[1] ZONDEK, B., *et al.*, *J. A. M. A.*, **128**, 939, 1945

[2] GALLI-MAININI, C., *Semana méd.*, **54**, 337 and 447, 1947; *Rev. Soc. argent. de biol.*, **23**, 125, 303, and 399, 1947; *J. Clin. Endocrinol.*, **7**, 653, 1947; *Endocrinology*, **43**, 349, 1948.

[3] ZWARENSTEIN, H., and D. G. DUNCAN, *Clin. Proc.*, **3**, 186, 1944.

[1] HAMBURGER, C., *Acta path. et microbiol. Scandinav.*, **18**, 457 and 485, 1941; *J. Endocrinol.*, **5**, xxiv (Proc.), 1947.

[2] OSTERGAARD, E., "Antigonadotropic substances," E. Munksgaard, Copenhagen, 1942.

[3] BROWN, W. E., and J. T. BRADBURY, *Am. J. Obst. & Gynec.*, **53**, 749, 1947.

migration of the testes into the scrotum. Injections of chorionic gonadotrophin (750 units, twice daily during 3 weeks) in cases of hypogonadism have been followed by stimulation of the Leydig cells, secretion of male hormone, and growth of the penis and sexual hairs; estrogens are excreted in the urine. The subsequent injection of FSH has provoked spermatogenesis in these cases.[1]

The hypophysis during pregnancy. The anterior hypophysis increases in size during pregnancy.[2] The so-called "pregnancy cells" appear in the pars distalis. These cells are seen not only in pregnancy but also after treatment with sexual hormones. According to Severinghaus, they are degranulated acidophil and basophil cells, loaded with mitochondria. The anterior hypophysis of pregnant women contains little gonadotrophin between the second and ninth months of pregnancy; the gonadotrophin content returns to normal after the delivery.

Hypophysectomy is followed by abortion in several species (dog, cat, and rabbit). In these species the secretion of the corpora lutea is indispensable for the persistence of pregnancy. Abortion can be prevented in hypophysectomized animals by treatment with progesterone, estrone, or gonadotrophins (Robson). In the rat nidation does not take place if the hypophysis is removed during the first days of pregnancy. If it is removed between the seventh and tenth day, the fetuses are reabsorbed. If it is removed from the eleventh to the twentieth day, the delivery is prolonged, the mother dies, and the fetuses are born dead.[3]

THE SEXUAL FUNCTIONS OF THE THYROID

The thyroid has an influence on sexual functions by means of a double mechanism: (*a*) it controls the secretion of gonadotrophin by the pars distalis, an adequate amount of thyroid secretion stimulating the secretion of gonadotrophin, while excessive amounts have as unfavorable an effect as thyroid insufficiency; (*b*) it acts directly on the ovaries.[4]

[1] Maddock, W. O., and W. O. Nelson, *J. Clin. Endocrinol.*, **12**, 985, 1952.

[2] Comte, 1898; Erdheim and Stumme, 1909.

[3] When considering the functions of the neurohypophysis facts were discussed which showed that it perhaps plays a part in delivery. The fundamental importance of the hypophysis in lactation will be discussed in Chap. 62.

[4] A direct action of the thyroid on the secondary sexual organs has not been demonstrated.

In the female the thyroid has a slightly greater weight than in males, and it varies in size and cytologic aspect in the different stages of sexual activity (puberty, pregnancy, menopause, and sometimes the menstrual cycle). Diseases of the thyroid are more frequently observed in women than in men, and the course of the disease varies with the stage of sexual activity of the patient.

Thyroid insufficiency prolongs the duration of diestrus and diminishes hypophyseal gonadotrophin in rats. In several species severe thyroid insufficiency at an early age causes hypoplasia of the sex glands and sexual insufficiency.[1] In women during the early stages of thyroid insufficiency there is an increase in menstrual flow; uterine hemorrhage and abortion frequently occur. During the later stages the menstrual flow diminishes; then amenorrhea and sterility are observed. In the male, severe thyroid insufficiency is sometimes accompanied by genital hypoplasia and loss of libido.

Mild hyperthyroidism provokes early puberty and a slight increase in the frequency of estrus in the rat. Severe hyperthyroidism causes retardation or suppression of the estrous cycle. Gonadotrophins in the hypophysis increase during the first stages (Reforzo) and diminish in advanced hyperthyroidism (Arrighi). The ovaries and sexual organs retrogress in severe hyperthyroidism and their response to gonadotrophins and sexual hormones is diminished.

In women suffering from hyperthyroidism, disturbances in the menstrual cycle are not always observed; usually as the condition increases in severity, menstruation and fertility diminish. Subtotal thyroidectomy improves the condition of the sexual functions together with the general improvement experienced by the patients.

Thyroid treatment often improves cases of hypomenorrhea, probably by its action on the anterior hypophysis; however, an excess of thyroid may cause amenorrhea.

THE SEXUAL FUNCTIONS OF THE ADRENALS

The adrenal cortex secretes hormones that exert a direct influence on the sexual characters. Probably corticoadrenal hormones also act on

[1] In the rabbit the follicles ripen, but ovulation is difficult to provoke; the anterior hypophysis contains FSH but does not have LH; injection of extract of such a hypophysis does not provoke ovulation (Chu).

the anterior hypophyseal secretion of gonadotrophins, but this has not been clearly demonstrated.

The adrenal cortex and the gonads arise from the epithelium of the celoma. The gonads have a definite effect on the adrenal cortex, as is shown by the differences between the cortex of the male and female. Thus in females the cortex is stimulated and has a greater development than in the male, in which androgenic hormones have a slight inhibitory influence. Castration is followed by moderate atrophy of the adrenal cortex in the female and slight hypertrophy in the male (Pinto, 1943). The adrenal of the female shows certain changes coinciding with estrus, sexual activity, and pregnancy.

The following substances with sexual activity have been extracted from the adrenal cortex: (*a*) female hormones (including estrone); (*b*) male hormones, such as adrenosterone and other steroids related to androsterone; (*c*) progestational hormones which act on the endometrium, such as progesterone.[1] Normally these adrenal steroids have little effect, and they are unable to prevent retrogression of the sexual organs following castration.[2] Estrogens and androgens found in the urine of castrates are supposed to be produced mainly by the adrenal cortex.

The masculinizing and feminizing effects of tumors or hyperplasia of the adrenal cortex have been considered in Chaps. 55 and 57.

[1] Survival after total adrenalectomy is more prolonged if the operation is performed during the animal's estrus or pregnancy, because in these states there are active corpora lutea.

[2] In young rodents atrophy of the prostate takes place more rapidly if adrenalectomy is performed at the same time as castration.

BIBLIOGRAPHY

American Medical Association, "Glandular Physiology and Therapy," Chicago, 1942.

Beach, F. A., "Hormones and Behavior," Hoeber, New York, 1948.

Beach, F. A., A Review of Physiological and Psychological Studies of Sexual Behavior in Mammals, *Physiol. Rev.*, **27**, 240, 1947.

Hamburger, G., Gonadotropins, in Emmens, C. W., "Hormone Assay," Academic Press, New York, 1950, p. 173.

Li, C. H., The Chemistry of Gonadotropic Hormones, *Vitamins & Hormones*, **7**, 223, 1949.

Li, C. H., and K. O. Pedersen, Physicochemical Characterization of Pituitary Follicle Stimulating Hormone, *J. Gen. Physiol.*, **35**, 629, 1952.

Martins, T., "Glandulas Sexuaes e Hypophyse Anterior," São Paulo, 1936.

New York Academy of Sciences, Protein Hormones of the Pituitary Body, *Ann. New York Acad. Sc.*, **43**, 253–421, 1943.

Venning, E. H., Assay of Hormone Secretions, *Methods in Med. Research*, **2**, 245, 1950.

White, A., The Chemistry and Physiology of Adenohypophyseal Luteotropin (Prolactin), *Vitamins & Hormones*, **7**, 254, 1949.

CHAPTER 59

The Endocrine Functions of the Ovary

Physiologic Anatomy. The surface of the ovary is covered by the germinal epithelium, formed by a layer of cells which are cuboidal in girls but become flat in adult women. In the course of fetal development the genital cords of Pflüger are formed by invagination of the germinal epithelium, later being divided into cell nests from which the primordial follicles arise (Fig. 269). In the newborn girl there is a large number of these primordial follicles (400,000), but with age they diminish, and they disappear completely after the menopause. Each primordial follicle consists of an oocyte surrounded by a single layer of follicular cells. The oocyte is converted by a process of ripening into the mature ovum; the other cells of the primordial follicle multiply and form the membrana granulosa, which is made up of several layers. Later, fluid appears between the cells of the granulosa (liquor folliculi), and a cavity, filled by this liquid,

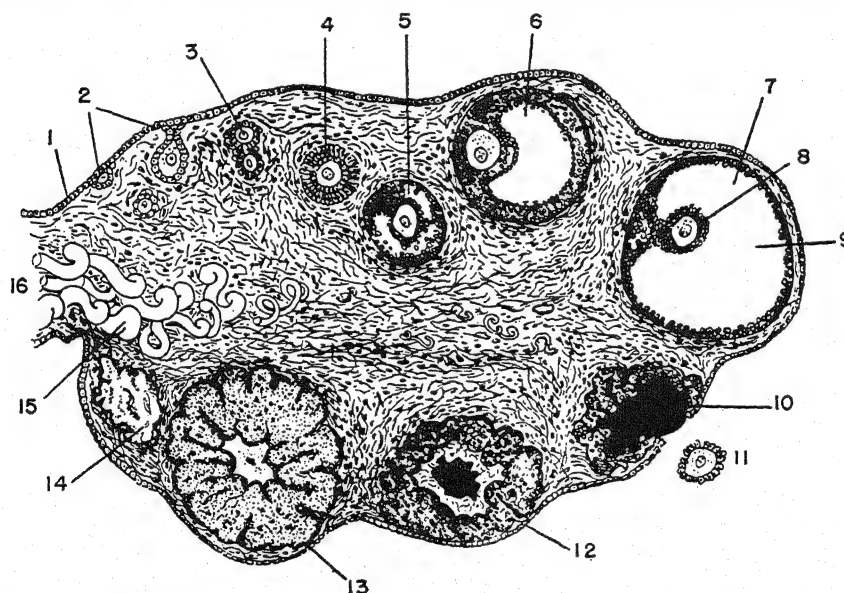



Fig. 269. Diagram of mammalian ovary, showing follicular maturation and the formation of the corpus luteum. 1, germinal epithelium; 2, ovigerous tubes; 3, egg nest; 4, double-layered follicle; 5, beginning of antrum formation in follicle; 6, follicle approaching maturity; 7, mature follicle; 8, ovum; 9, antrum filled with liquor folliculi; 10, corpus hemorrhagicum, *i.e.*, ruptured follicle filled with blood; 11, discharged ovum surrounded by cells from the granulosa; 12, young corpus luteum; 13, fully developed corpus luteum; 14, corpus albicans; 15, blood vessels; 16, mesovarium. (*After Patten, "Embryology of the Pig," The Blakiston Company, Philadelphia, 1927.*)

FIG. 270. Ovarian follicle. 1, germinal epithelium; 2, stigma; 3, granulosa; 4, discus proligerus; 5, basal membrane; 6, tunica albuginea; 7, internal theca; 8, external theca.

gradually appears. At one end of the follicle, a mass of cells surrounds the ovum; this is the discus proligerus, which forms a salient into the follicular cavity. The follicle is surrounded by a capsule formed by cells derived from the neighboring stroma. This capsule has two layers. The inner layer, called the theca interna, is made up of large polyhedral cells with an oval nucleus; it has many blood vessels, whence the name "tunica vasculosa" that is sometimes given it. The outer layer, or theca externa, is made up of fusiform cells and thick connective fibers, closely packed in concentric order; it is sometimes called the "tunica fibrosa" (Fig. 270).

FIG. 271. Human ovary. Two follicles and a corpus luteum. (*Courtesy of Professor R. Sammartino.*)

The mature graafian follicle is ruptured at the moment of ovulation, and the ovum, surrounded by cells of the discus proligerus and follicular fluid, is expelled into the peritoneal cavity and conveyed along the fallopian tube. Only one follicle ripens and expels an ovum at each monthly cycle (exceptionally, two); therefore only 300 or 400 follicles reach full maturity throughout the whole sexual life of a woman. The remaining hundreds of thousands of follicles disappear by a process of degenerative involution called "atresia," which can occur at any stage in the process of maturation.

The ruptured follicle is converted into the corpus luteum (Fig. 271). The follicular cavity is at first filled by erythrocytes, which have migrated by diapedesis from the blood capillaries, or as the result of a

minute hemorrhage. The cells of the granulosa rapidly multiply and form folds which fill the whole follicular cavity. These cells are filled with a lipoid pigment, which is either yellow or orange or white according to the species. The pigment is called "lutein" and the cells containing it are known as "luteal" cells. Connective tissue and numerous blood capillaries are found between the large luteal cells (Fig. 272). The cells of the theca interna take on the same aspect as the luteal cells in some species (theca-luteal cells). Later the corpus luteum undergoes a process of involution. As this process advances, neutral fats in the luteal cells increase, the organ takes on a bright yellow color, and a hyaline substance appears between the luteal cells, which are finally replaced by connective-tissue cells. The whole organ atrophies (corpus albicans) and eventually is reduced to the fibrous tissue of a residual scar. All corpora lutea pass through the following stages: (*a*) proliferation; (*b*) vascularization; (*c*) full development; (*d*) involution.

There are several types of corpora lutea, according to the circumstances of their development. In the following types the corpora lutea are of short duration: (*a*) the *periodic* or *cyclical* corpus luteum follows ovulation at each cycle; (*b*) the *atretic* corpus luteum is the result of the luteinization of the cells of a follicle that has not ripened completely and retains the ovum.[1] The corpus luteum has a longer period of activity in the following circumstances: (*a*) in *pregnancy*, throughout which the corpus luteum can be found, although from the fifth month onward it shows signs of involution; (*b*) in *lactation, i.e.*, as long as the mother nurses the child; (*c*) in *pseudopregnancy;* (*d*) after *hysterectomy* in certain species.

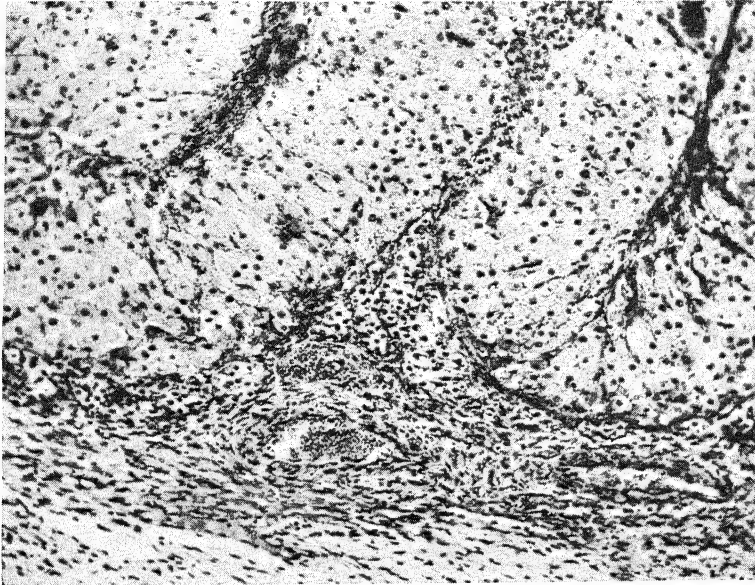



FIG. 272. Corpus luteum recently formed, surrounded by internal theca cells. (*Courtesy of Professor R. Sammartino.*)

The interstitial cells of the ovary form nests or columns; they have the aspect of epithelial cells, with granules of lipids and other substances. They are derived from the fusiform stroma cells and from the cells of the membrane of atretic follicles. They reach maximum development in girlhood; in adult women they are reduced to a few irregular columns of epithelioid cells in the midst of the stroma, or may even be completely missing. In the rabbit and other rodents, a considerable number of interstitial cells are found at all ages.

FUNCTIONS OF THE OVARY

The following are the principal functions of the ovary: (*a*) *reproduction* or *gametogenesis*, *i.e.*, the production of ova; (*b*) *endocrine activity*, *i.e.*, the secretion of female hormones which develop and maintain the secondary sexual characters; (*c*) the *formation of the corpus luteum*, which also secretes hormones; (*d*) the *preparation of the uterus for pregnancy* and the continuance of pregnancy;

[1] This type of corpus luteum is produced by the luteinizing hormone of the hypophysis and by chorionic gonadotrophins.

(*e*) the *development of the mammary gland* in preparation for lactation.[1]

REGULATION OF THE ENDOCRINE FUNCTIONS OF THE OVARY

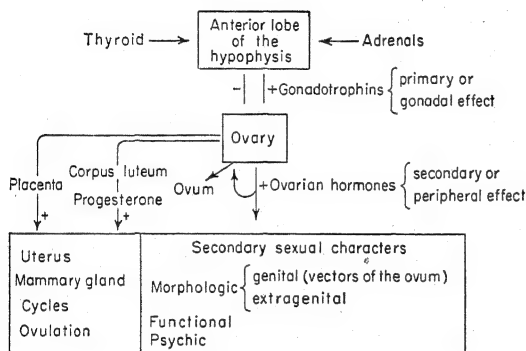
The development and functions of the ovary depend on several factors. The principal factor is the secretion of gonadotrophins by the pars distalis of the hypophysis, which controls the development, structure, and activity of the ovary (see diagram).

Thyroid → Anterior lobe of the hypophysis ← Adrenals

− | +Gonadotrophins { primary or gonadal effect

Ovary

Placenta | Corpus luteum | Progesterone | Ovum | +Ovarian hormones { secondary or peripheral effect

Uterus Mammary gland Cycles Ovulation	Secondary sexual characters Morphologic { genital (vectors of the ovum), extragenital Functional Psychic

The sexual function of the anterior lobe of the hypophysis is controlled by (*a*) *hormones* of the ovary, thyroid, and adrenal cortex; (*b*) *reflexes* stimulated by copulation, light, etc.; (*c*) *nutritive factors*, such as vitamin E, undernourishment, etc.

Hormones secreted by the ovary not only have a feminizing effect, they also act on the anterior hypophysis: (*a*) they moderate the secretion of gonadotrophins; (*b*) in certain concentrations they provoke the release of luteinizing hormone, which produces the rupture of the mature follicles (ovulation) with subsequent luteinization of these follicles and sometimes of other immature follicles (atretic luteinization).

Ovarian hormones act on the ovary itself, mainly indirectly through the hypophysis, but also directly on the follicles (see "Hormonal factors in the menstrual cycle").[2]

Factors that modify ovarian functions. The greater part of the factors that modify ovarian functions act through the anterior hypophysis, but some of them act directly on the ovary. The activity of the ovary is modified by extraovarian factors in the following circumstances:

1. In certain species ovulation is provoked by a reflex (copulation in the cat and rabbit, amplexus in the toad, visual reflexes in the pigeon) or by electrical stimulation of the hypothalamus or the whole brain (rat and rabbit).
2. Excessive illumination provokes sexual activity during hibernation in certain species.
3. Cold inhibits ovarian activity and suppresses the sexual cycle.
4. Infections and other diseases (*e.g.*, tuberculosis) may suppress the menstrual cycle.
5. Chronic undernourishment diminishes and eventually suppresses the sexual cycle and ovulation.
6. Thyroidectomy or thyroid insufficiency prevents sexual development and diminishes the frequency of the cycles. Moderate hyperthyroidism increases the frequency of the cycles, but severe hyperthyroidism suppresses the cycle.
7. Adrenal insufficiency and corticoadrenal hyperfunction disturb ovarian functions in women.

There is an optimum level of thyroid and of adrenal secretion necessary for the normal functioning of the ovary; insufficiency or excess of these hormones disturbs the activity of the ovary.

THE SEXUAL CYCLE

Comparative biology of the sexual cycle. Cyclic activity is an outstanding feature of feminine sexual functions. The sexual cycles begin at puberty and end at the menopause. In some species there are only one or two periods of sexual activity in the year, *i.e.*, one cycle at each breeding season. These periods during which the female accepts the male are called *estrus* or "heat"; outside estrus, *i.e.*, during *anestrus*, the female does not accept the male. Animals with this type of cycle are called *monoestrous*. At the beginning of the cycle, in the stage known as *proestrus*, there are congestion and secretion in the uterus and vagina, the vaginal epithelium proliferates, and the vulva is swollen. During estrus the changes in the genital tract are at a maximum, ovulation takes place, the vaginal epithelium desquamates cornified cells, and the female accepts the male. The stage following

[1] The part played by the ovary in the development and maintenance of female secondary sexual characters has already been considered in previous chapters.

[2] The role of the thyroid and adrenals in the regulation of anterior hypophyseal sexual functions, of ovarian functions, and of secondary sexual characters has been discussed in preceding chapters.

estrus, during which the corpus luteum is formed and secretes, is called *metaestrus*. In some species (*e.g.*, the rabbit and ferret), even when the female has not been fertilized, estrus is followed by changes similar to those of pregnancy, which last as long or nearly as long as pregnancy. There is considerable development of the mammary glands and sexual organs and changes in the uterine mucosa which resemble those seen just before nidation of the embryo (progestational changes). This condition is called "pseudopregnancy."

In *polyestrous* animals the cycles occur with greater frequency throughout the sexual life of the female. Each cycle has the proestrus, estrus, and metaestrus phases and is separated from the following cycle by a short period of quiescence called *diestrus*. Rats and mice are polyestrous animals. Their sexual cycle has been the subject of considerable work, on which a great part of the knowledge of female sexual functions is based. There are changes in the ovary, fallopian tube, uterus, and vagina. The vaginal response is typical and can be easily followed by microscopic examination of smears of the vaginal content.[1] During proestrus large nucleated epithelial cells are found; during estrus there are only squamous cornified cells; during metaestrus leukocytes appear among the cornified cells; during diestrus there are only leukocytes, a few epithelial cells, and mucus. The corpus luteum has a very short time of activity in polyestrous animals, such as the rat and mouse; luteal activity is much more intense and prolonged in species in which estrus is followed by pseudopregnancy.

The average duration of the cycle in different species is as follows: rat and mouse, 4 to 6 days; guinea pig, 14 days; sheep, 16 days; cow, 20 days; sow, 21 days; macaque, 27 days; woman, 28 days; chimpanzee, 36 days. The bitch has two cycles a year, the bat and marmot one cycle a year. In the cat, rabbit, and ferret the follicle ripens but ovulation does not take place unless there is copulation or intense sexual excitement. The rabbit ovulates 12 to 16 hr. after copulation and the cat 26 hr. after.

The Sexual Cycle in Women

Puberty. The most typical sign of female puberty is the onset of menstruation. In temperate climates this usually occurs between the ages of eleven and fifteen (average, thirteen and one-half); if it takes place outside these ages, puberty is considered as precocious or retarded.[1] Other signs of puberty are the development of the ovaries, uterus, and mammary glands, the appearance of hair on the pubis and in the axillae, etc.

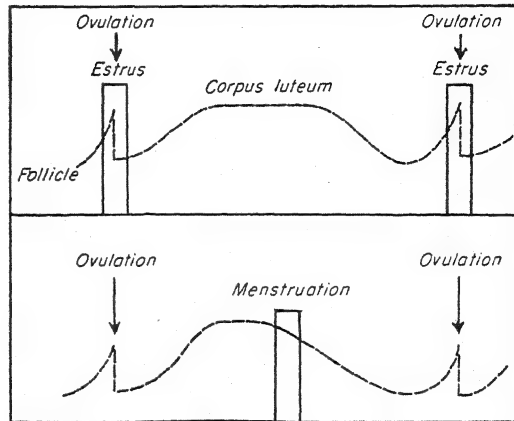



Fig. 273. Diagram of sexual cycle. Above, in species with estrus; below, in primates.

Menstruation. The two outstanding events in the female cycle are menstruation and ovulation, the latter occurring toward the middle of the period between two menstruations (Fig. 273). Ovulation does not always occur at each cycle. In early puberty there are many anovulatory cycles. In some women they also occur from time to time, with greater or lesser frequency. These anovulatory cycles are sterile, because there is no ovum to be fertilized.

Menstruation consists in a hemorrhagic flow from the vagina (20 to 200 cc.), which comes from the endometrium. Menstrual blood usually does not clot. This flow lasts 3 to 7 days in 95 per cent of women (3 to 5 days in 66 per cent). It appears every 27 or 28 days, but there is considerable variation in the duration of the menstrual cycle of different women (Fig. 274), and even in the same woman the duration of the cycle may change from time to time. In the course of the 30 or 40 years of sexual life, a woman menstruates from 300 to 500 times.

Ovulation. Ovulation takes place toward the middle of the cycle. Ova have been collected by washing out the fallopian tubes in women be-

[1] Stockhard, C., and C. Papanicolau, *J. Anat.*, 22, 225, 1917.

[1] In well-nourished girls menstruation begins earlier than in the undernourished.

tween the twelfth and twenty-first day after the beginning of menstruation; the largest number have been found on the fifteenth day (Allen). The postovulatory stage, *i.e.*, the interval between ovulation and the following menstruation, is less variable in duration (10 to 16 days) than the preovulatory stage, *i.e.*, the interval between the beginning of menstruation and ovulation. Since the ova do not live long, fertilization can occur only during a short period following ovulation; this is the fertile period of the cycle. According to Ogino, Knaus, and others, ovulation occurs 12 to 16 days before the following menstruation. The situation of the fertile period in the cycle varies, therefore, with the length of the cycle. It can fall from the eighth to the twenty-first day (Hartman); the days before and after this period are sterile in the sense that fertilization cannot take place during them. In female monkeys the existence of fertile and sterile periods has been clearly demonstrated (Hartman). Observations in women have shown increased fertility in the middle of the cycle, but fertilization has been seen to take place even during the so-called "sterile" period.

FIG. 274. Frequency distribution of duration of menstrual cycle. Abscissa, days of cycle; ordinate, percentage of incidence in 154 women. (*After Haman.*)

Diagnosis of ovulation. Ovulation has been observed in women by inspection of the ovaries in the course of abdominal operations. Definite proof of ovulation has been obtained by washing out the fallopian tubes and observing the ovum. There are, however, more simple and practical methods of establishing the moment of ovulation. Basal temperature (rectal or sublingual), taken in the morning, varies between 36.3°C. and 36.8°C. during the preovulatory phase and rises to 37 to 37.3°C. in the postovulatory phase. This difference of 0.3 to 0.5°C. is observed only in ovulatory cycles, not in anovulatory ones (Fig. 275).

At the moment of ovulation there is an increase in the cornified acidophil cells of the vaginal smear (Allende and Orías) (Fig. 276). There is also an increase in urinary gonado trophin just before ovulation occurs,[1] which has been used for diagnostic purposes. Two rats, 21 to 25 days old, are injected subcutaneously with 2 cc. of urine, and 2 hr. later congestion of the ovaries is observed if the gonadotrophin content of the urine is high. The test is repeated for four days, and if it is consistently positive, ovulation will take place on the fourth day, this being the most favorable moment for fertilization. It usually occurs 11 to 14 days after the

FIG. 275. Temperature changes during menstrual cycle. Abscissa, days of cycle; ordinate, temperature in degrees centigrade; stippled area, menstruation.

[1] CORNER, G. W., E. J. FARRIS, and G. W. CORNER, JR., *Am. J. Obst. & Gynec.*, **59**, 514, 1950.

first day of the last menstruation, or at the half of the cycle minus 2 days (Farris).

In an ovulatory cycle a corpus luteum is formed, and its activity can be established by (*a*) glandular proliferation and secretion of the endometrium, established by histological examination of a small fragment of uterine mucosa excised with a curet; (*b*) a sudden increase in the urinary excretion of pregnandiol (a metabolic product of progesterone) for 4 to 5 days following ovulation; (*c*) an increase of 0.3 to 0.5°C. (0.6 to 0.9°F.) in basal temperature.

Hormonal factors in the menstrual cycle. When the follicle begins to ripen, the oocyte and the cells of the granulosa stimulate growth and activity of the cells of the theca interna. In some of the follicles the theca secretes estrogens, which provoke secretion of follicular fluid by the cells of the granulosa and sensitize the follicle to FSH secreted by the hypophysis. In women FSH provokes the maturation of only one follicle (in rare cases, two) at each cycle, *i.e.*, the follicle sensitized by estrogens. This follicle develops up to a diameter of 2 cm., while the other follicles that were in the process of ripening cease to grow, and some become atretic. The mass of maturing follicles is nevertheless necessary for the development of the single follicle that ovulates.[1]

Secretion of estrogens increases gradually, reaching a maximum at the moment of ovulation. During the preovulatory phase estrogens predominate and stimulate the development of the uterus, especially the proliferation of the endometrium. When estrogens are at a high level, secretion of FSH is inhibited and secretion of LH, and probably of luteotrophin (prolactin), increases. LH acting on the mature and sensitized follicle causes its rupture and the release of the ovum, *i.e.*, ovulation. LH then provokes the development of the corpus luteum in the ruptured follicle, and luteotrophin stimulates the secretion and maintains the structure of the newly formed corpus luteum.

Progesterone, secreted by the corpus luteum, stimulates secretion of the endometrial glands and causes progestational changes which prepare the endometrium for the implantation of the fertilized ovum. During the postovulatory phase progesterone activity is predominant. Three to four days after ovulation pregnandiol appears in the urine; from 40 to 60 mg. of this metabolic product of progesterone are excreted during each cycle (Venning and Browne, 1936). The high level of progesterone in the circulation inhibits the secretion of LH and luteotrophin, and this causes involution of the corpus luteum. Progesterone then diminishes, the endometrium breaks down, and menstruation occurs.

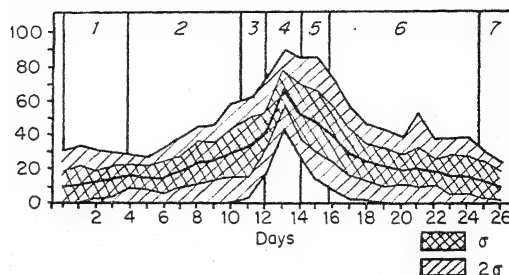



FIG. 276. Cornification of vaginal epithelium in the course of a cycle in which ovulation occurs. 1, menstrual; 2, postmenstrual; 3, preovulatory, 4, ovulatory; 5, postovulatory; 6, progestational; 7, premenstrual. (*I. C. de Allende and O. Orías, 1947.*)

Apparently normal menstruation with typical changes in the endometrium can be provoked in cases of amenorrhea and in castrated women by injecting estrogens followed by progesterone; when progesterone is withdrawn, uterine bleeding commences. Uterine hemorrhage has also been provoked in castrated female monkeys and in castrated women by withdrawal of treatment when estrogens only are injected (Corner), or even by reducing the dose of estrogen (Zuckerman). These facts led Corner to formulate the hormone-withdrawal hypothesis of menstruation. According to this hypothesis, menstruation is caused by the sudden fall in progesterone secretion due to involution of the corpus luteum in the final stages of the cycle.

Anovulatory cycles, *i.e.*, menstrual cycles without ovulation, have been observed in monkeys and in women. These cycles are sterile; fertilization cannot take place because no ovum has been released. No corpus luteum develops in this type of cycle; therefore the endometrium does not show progestational changes or the secretory phase. According to Corner's hypothesis, menstruation in these cycles is caused by a decrease in estrogens.

The uterine cycle. Typical changes take place in the endometrium in the course of the

[1] The factors conditioning the favored follicle so that "it is lifted on the shoulders of its contemporaries" are not yet well understood (Hohlweg and Westman, 1934; Laqueur *et al.*, 1945–1946; etc.).

menstrual cycle (Fig. 277). These have been studied by simultaneous microscopic observation of small fragments of vaginal mucosa and endometrium, correlating the changes seen with those observed in the ovary in women submitted to abdominal operations. Simple macroscopic examination of the ovaries *in situ* may be the cause of serious errors (Hartman); therefore histological examination should always be performed. The following changes occur in the course of a complete cycle:

1. *Proliferative stage* (sixth to fourteenth day after the beginning of menstruation). The endometrium increases in thickness, and the glands lengthen. At first they are straight, but later they become folded and tortuous (Figs. 277*a* and 278). This stage is sometimes called the "estrogenic phase" because estrogens stimulate the growth of the endometrium.
2. *Secretory or progestational stage* (fifteenth to twenty-seventh day of cycle). The endo-

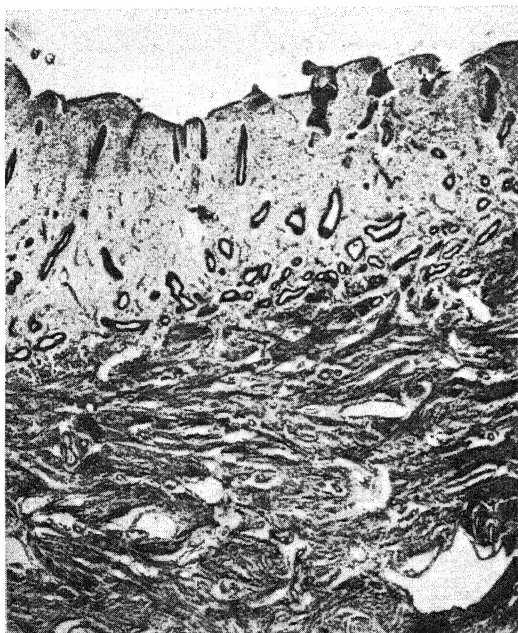
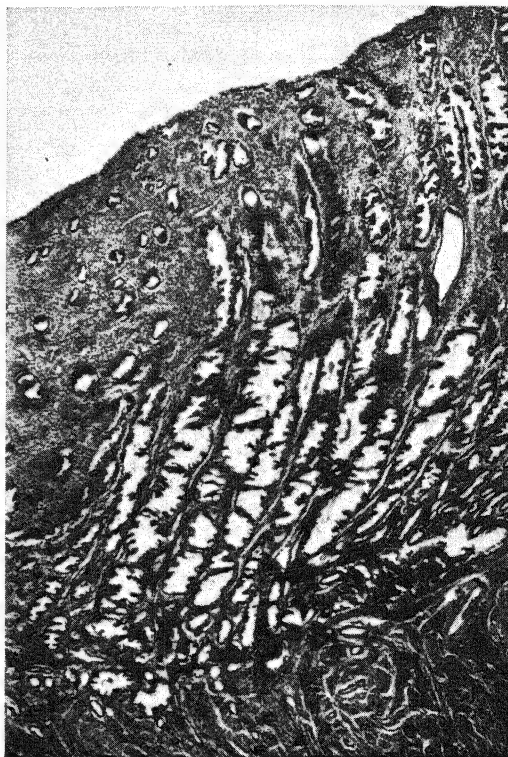
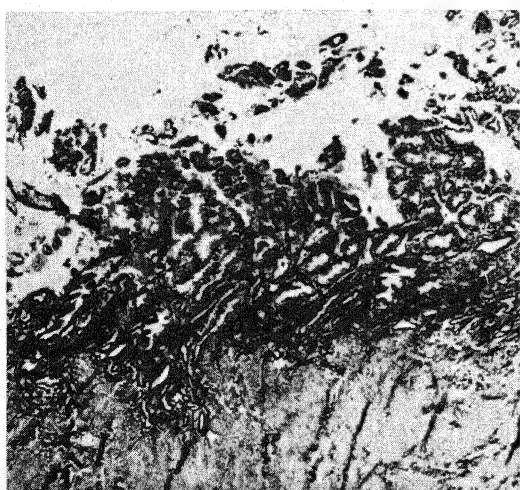
a

b

c

FIG. 277. Cyclical changes in human endometrium. *a*, proliferation; *b*, secretion; *c*, menstrual desquamation.

metrium increases in thickness even more. The cells of the stroma proliferate. The uterine glands attain a considerable development and are distended by mucus. On section the endometrium has a serrated aspect owing to the epithelial folds. The spiral arteries in-

FIG. 278. Hormonal factors in human female sexual cycle. FSH, follicle-stimulating hormone; LH, luteinizing hormone; LTH, luteotrophic hormone. Above: 1, primary follicle; 2, ripening follicle; 3, ripe follicle; 4, rupture of follicle (ovulation); 5, corpus luteum; 6, developed corpus luteum; 7, degenerating corpus luteum (no pregnancy). Endometrium proliferates during the first phase of the cycle (estrogen) and secretes during the second phase (progesterone). Below: Urinary elimination of estrogen in international units in 24 hr. (left), pregnandiol in milligrams (right), and gonadotrophins in international units per 24 hr. (extreme right). (*Diagram below, courtesy of Dr. Eleanor H. Venning, Royal Victoria Hospital, Montreal.*)

crease in size and number, and there are numerous wide capillaries (Figs. 277*b* and 278). Changes in this stage are produced by the hormone of the corpus luteum (progesterone). They are a preparation for the reception and implantation of the fertilized ovum; hence the name "progestational stage."

3. *Destructive stage.* The superficial spongelike layers of the endometrium diminish in thickness, the circulation becomes slow, and stasis occurs. Later there is vasoconstriction in the mucosa which lasts 4 to 24 hr., a fact that has been directly observed in transplants of endometrium placed in the eye behind the cornea.[1] The arteries again dilate, and hemorrhage begins. Blood is lost through the capillaries, arterioles, and veins of the stroma and glands. Small blood lakes are formed

[1] MARKEE, J. E., "Progress in Gynecology," Grune & Stratton, New York, 1946; REYNOLDS, S. R. M., *J. A. M. A.*, **135**, 552, 1947.

under the superficial epithelium, which becomes necrosed and is sloughed into the uterine cavity together with unclotted blood, leukocytes, and mucus (Figs. 277*c* and 278). If there is abundant hemorrhage, thrombin and blood clots are formed; if there is a slow hemorrhage, the blood is gradually defibrinated and contains antithrombin. Fibrin in intrauterine clots is usually dissolved fairly rapidly (fibrinolysis).

In approximately three days the endometrium loses half its thickness. The process of repair then begins (second to fifth day) and merges into the proliferative stage.

A menstrual toxin (menotoxin) has been described,[1] which seems to be an atypical euglobulin produced by the endometrium, similar to necrosin, found by Menkin in inflammatory exudations. This substance has been found not only in menstrual blood but also in the blood of the general circulation. It is believed to cause vascular lesions. Release of this toxin seems to be due to the involution of the corpus luteum and to the decrease in estrogen and progesterone.

The significance of menstruation has been the subject of much discussion. It is not the equivalent of estrus, to which the preovulatory stage could be more appropriately compared (Fig. 273). Sometimes there is a small hemorrhage and pain (*Mittelschmerz*) toward the end of the proliferative stage which corresponds, to a certain extent, to the phenomenon of estrus. In women there is no typical estrus and the male is accepted at any stage of the cycle. Menstruation provokes renewal of the endometrium when the progestational changes brought about in preparation for the implantation of the fertilized ovum are not followed by pregnancy.

"Amenorrhea" means the cessation of menstruation. "Hypomenorrhea" means a scanty menstrual flow, and "hypermenorrhea" an abundant flow. "Oligomenorrhea" is the term used for the appearance of menstruation at long intervals, and "polymenorrhea" for abnormally frequent menstruation. "Dysmenorrhea" means painful menstruation.

The vaginal cycle. (Figs. 276 and 279.) In childhood, in old age, and in genital aplasia, *i.e.*, whenever there is estrogen deficiency, the vaginal epithelium is thin. Estrogens provoke an increase in the thickness of the epithelium and cornification of the superficial cells. The aspect of the vaginal smear changes with the level of estrogenic activity: (*a*) estrogenic activity is revealed by the preponderance of superficial karyopycnotic and cornified acidophile cells; (*b*) when there is slight or moderate estrogen deficiency, the majority of the cells are small deep-layer cells and intermediate cells, and a few superficial cells, with mucus, leukocytes, and bacteria; (*c*) in marked estrogen deficiency, there are only small deep-layer cells, mucus, leukocytes, and bacteria. Cornified cells increase as the cycle proceeds, remaining separated from each other. They are at a maximum (peak of cornification) at the moment of ovulation. During the progestational stage there is a preponderance of intermediate cells; cornified cells are seen in clumps with folded margins.

Cytological examination of the urethral smear or the urinary sediment shows changes similar to those of the vaginal smear.[1]

The vaginal pH is lower during the first half of the cycle than during the second half.

Changes in other organs. Cyclic changes are also observed in other epithelia, such as the mucosae of the fallopian tubes, the mouth, the nose, and perhaps the mammary gland. In several species cyclic changes in the adrenals have been reported.

General symptoms. Just before menstruation, several disturbances, known as "premenstrual symptoms," are frequently observed. These include headache; tension in the pelvis, which is sometimes intensely painful; and slight or even marked congestion of the nasal, laryngeal, and buccal mucosae. This congestion sometimes causes slight hemorrhage, which has been mistakenly called "vicarious menstruation." Occasionally tension and tenderness of breasts are noted. In some women water is retained and there is a transitory increase in weight. In the female monkey estrogens (secreted by the ovary or injected) provoke congestion and edema of the sexual skin, and the body weight may increase by 25 per cent. When estrogenic activity ceases (ovulation, or withdrawal of estrogen treatment), congestion and edema of the sexual skin disappear, and there is polyuria and loss of weight, before menstruation begins. During menstruation the BMR and temperature diminish. (Deep rectal or vaginal temperature is

[1] Smith, O. W., and G. van S. Smith, *J. Clin. Endocrinol.*, **7**, 483, 1946; *Am. J. Obst. & Gynec.*, **54**, 201 and **212**, 1947.

[1] Del Castillo, E. B., J. Argonz, and C. Galli-Mainini, *J. Clin. Endocrinol.*, **9**, 1362, 1949.

low during the first half of the cycle; there is a further sharp decrease at the moment of ovulation and then a rise, and the temperature is high during the activity of the corpus luteum.)

upper part of the chest, together with a sensation of suffocation and sweating. Frequently a tendency to obesity appears and sometimes there are signs of virilization (hair growth on the upper lip and on the body, and lowering of the voice register).[1] Personality changes often occur at the menopause; there are emotional instability, psychic tension, and irritability. Some women complain of headache and of diffuse pain in the trunk and limbs.

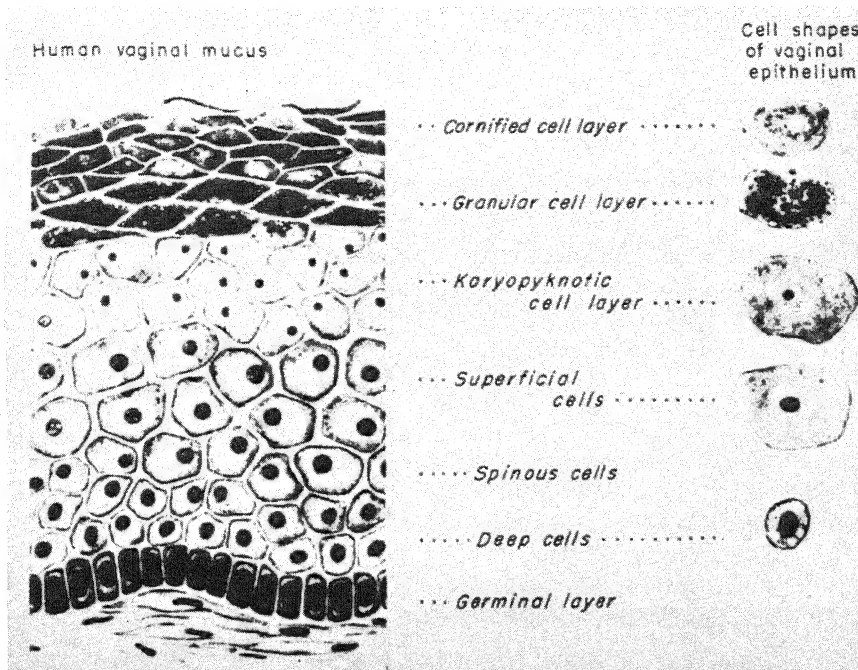



FIG. 279. Diagram of vaginal cycle in women. (*After Murray.*)

Menopause. The menopause, or climacteric, consists in the cessation of the menstrual cycle.[1] It occurs most frequently between forty-five and fifty years of age. If it takes place before the age of thirty-five it is considered premature, and retarded if the cycles are prolonged beyond the fifty-fifth year. It is known as "artificial menopause" when it is due to castration by surgical procedures or by the application of x-rays, or to hysterectomy. Natural menopause is accompanied by a decrease of, and finally the disappearance of, estrogens in the urine and by an increase in gonadotrophins.[2] Ovarian activity diminishes and finally ovulation ceases. The genital organs and mammary glands undergo gradual involution. There are many general symptoms, including crises of vasodilatation (hot flashes) in the skin of the head, neck, and

[1] MARAÑON, G., "La Edad Crítica," 2d ed., Ruiz, Madrid, 1925.

[2] Many of the unpleasant symptoms suffered by women at the menopause have been attributed to the increase in gonadotrophin.

Castration. The removal of one ovary does not provoke the appearance of symptoms of ovarian insufficiency or disturbances in the cycle, nor does it diminish the number of offspring if the remaining ovary is in normal condition. Extirpation of both ovaries is followed by symptoms of ovarian insufficiency. If castration is performed before puberty the genital organs and the secondary sexual characters do not develop. There are few reported cases of prepuberal castration in women, and most of these have not been the object of a thorough and complete study. The genital tract and mammary gland remain in an infantile condition. There is no menstruation. Axillary

[1] This has been attributed to androgenic corticoadrenal hyperfunction.

and pubic hair does not grow. Feminine and genital instincts do not appear. The subjects are usually tall owing to retarded closure of the epiphyseal cartilages. The condition is called "sexual infantilism" ("feminine eunuchoidism" is not a correct term), and it has been observed in different degrees of severity.

Postpuberal castration is performed for therapeutic reasons in adult women suffering from certain diseases of the ovaries or uterus. In this case the disturbances observed differ from those of sexual infantilism because secondary sexual characters are already developed. The disturbances are similar to those seen in the menopause, but they are more severe than in the natural menopause and take place more rapidly.[1] There is a more or less intense and progressive atrophy of the uterus and the epithelia of the vagina and vulva. Menstruation ceases, but in some cases castration is followed by one menstrual flow. The mammary gland atrophies, although in some cases this is not apparent because of the increase in adipose tissue, due to a tendency to obesity. The libido frequently suffers no change after castration in women, because it is strongly conditioned by psychic factors, but in animals estrus ceases completely. Growth of hair on the face and limbs often occurs, and sometimes hypertrichosis is seen. This has been attributed to an increase in corticoadrenal androgenic secretion, but it might also be due to the suppression of an inhibitory effect of estrogens on hair growth, similar to the inhibitory effect on the spurs of the hen. Castration is often followed by psychic disturbances such as emotional instability, irritability, or depression; the subjects become very susceptible to emotional stimulation and are easily angered or subject to fits of uncontrolled laughter. They also suffer from vasomotor disturbances (hot flashes) accompanied by sweating; these flashes can be provoked by the injection of adrenaline. They often complain of palpitations (tacnycardia is observed), cold hands and feet, headache, diffuse pains in the limbs, and other abnormal sensations. Urinary elimination of estrogen diminishes and the excretion of gonadotrophins increases, especially during the vasomotor crises.

[1] The examination of vaginal smears shows that at the menopause the ovary continues to secrete estrogens at a progressively decreasing level during several years. Castration causes a sudden cessation of estrogen secretion.

Restitution of ovarian functions. The effects of castration can be prevented or controlled by several methods (see Chap. 57): (*a*) ovarian grafts; (*b*) implantation of ovarian tissue; (*c*) injection of ovarian extract; (*d*) injection or implantation (pellets) of ovarian hormones. The most commonly used method is the administration of ovarian hormones (natural estrogens) or of substances that have estrogenic activity. This treatment stimulates the development of the genital tract and maintains it as long as the treatment lasts. In young castrated women, if the uterus remains intact, menstrual cycles can be provoked by the injection of 15 to 20 mg. of estradiol during 20 days, followed by 5 to 25 mg. progesterone during 5 days; when the treatment is suspended the menstrual flow begins. Biopsies of the endometrium show the histologic changes corresponding to the three stages of the cycle.

Estrogens.[1] The knowledge of the cyclic changes occurring in the vaginal epithelium of rodents provided a simple method for the study of substances with estrogenic activity. At first the natural estrogens, *i.e.*, those obtained from animal tissues and fluids, were studied.[2] Later artificial estrogens were prepared. Dozens of estrogenic substances are now known. They have been obtained from animal and plant tissues and from minerals such as tar and asphalt.

Natural estrogens are steroids that have a perhydrocyclopentenephenantrene ring. They are derivatives of the unsaturated Δ, 1, 3, 5-estratriene ring. The principal estrogenic substances[3] are: (*a*) *estradiol;*[4] (*b*) *estrone;*[5] (*c*) *estriol.*[6] These three substances have an OH group on C^3; estradiol also has OH on C^{17}, estrone O on C^{17}, and estriol OH on C^{16} and C^{17}. Estradiol has been obtained from the ovaries of sows, from human urine, and from the testes of stallions. It is six times as active as estrone and seems to be the main, if not the true, ovarian hormone. In therapeutics the benzoate, propionate, or dipropionate of estradiol is used.

[1] Doisy, E. A., *J. A. M. A.*, **116**, 501, 1941; *Endocrinology*, **30**, 933, 1942.

[2] Mainly by Doisy, Butenandt, and Laqueur and their associates in 1929 and 1930.

[3] The names "folliculin" and "estrin" are no longer used.

[4] Dihydroxyestrin.

[5] Ketohydroxyestrin or theelin.

[6] Trihydroxyestrin or theelol; it is found in the placenta forming part of a complex (emmenin).

Natural estrogens have been found in the ovary, liquor folliculi, placenta, adrenals, and the human corpus luteum during the first third of pregnancy. There are large quantities in the urine of pregnant women and mares and in stallion's urine[1] (Table 78).

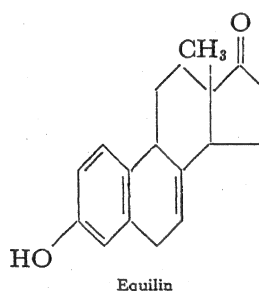
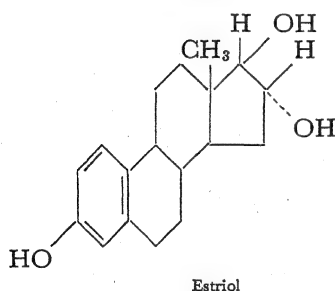
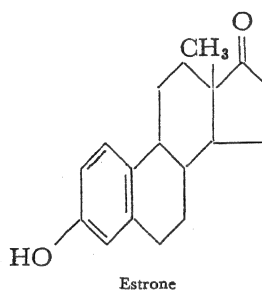
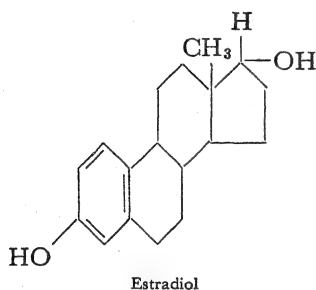
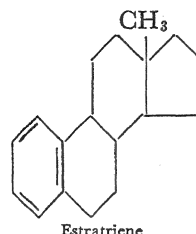
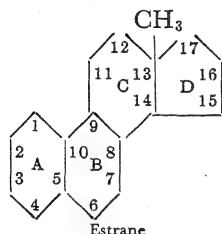
Estrane

Estratriene

Estradiol

Estrone

Estriol

Equilin

Castration diminishes the urinary output of estrogens in the horse. The urine of normal women contains little estrogen; there is an increase about the time of ovulation and another just before menstruation (Fig. 278).

Ovariectomy performed during the first weeks of pregnancy does not prevent the increase in the urinary excretion of estrogens, which occurs later and is maintained at the usual high level in spite of castration. This is due to the fact that estrogens in pregnant women's urine are produced not by the ovaries but by the placenta.

The cells of the granulosa, the theca interna, and the corpus luteum all seem to produce estrogens, and when a large number of follicles ripen there is hypersecretion of estrogen. The follicles, however, are not the only structures that can produce estrogens, because normal cycles have been observed in rodents in which the follicles were destroyed by x-rays without damaging the interstitial cells. The latter therefore can also produce estrogens.

[1] Large quantities of estrogens are found in the urine of pregnant women, pregnant mares and stallions. In women most of the estrogen is estriol.

Table 78. Estrogen Content of Tissues and Organic Fluids

Estrogen	*Sow's ovary, mg. per kg.*	*Pregnant woman's urine, mg. in 24 hr.*	*Human placenta, mg. per kg.*	*Bovine adrenal*	*Stallion's urine*
Estradiol. . .	0.14	0–0.06	0.038		
Estrone. . . .	0.10	1–3	0.033	+	+
Estriol.	. . . .	10–30	0.140		

Estrogens produced by the organism are rapidly converted into inactive substances; only small amounts are eliminated without change. The most important site of estrogen inactivation is the liver. It is for this reason that, to obtain the effects of a certain amount of estrogen injected subcutaneously, fifty to one hundred times that amount must be given by mouth.[1]

Diethylstilbestrol

Hexestrol

Dienestrol

The most commonly used artificial estrogens are diethylstilbestrol, hexestrol, dienestrol, bencestrol, ethylnilestradiol, doysinolic and dehydrodoysinolic acids, allenolic and dimethylallenolic acids. The first three are derivatives of dihydroxystilbene (stilbestrol) and they differ from each other by their side chains. They have the same estrogenic properties as natural estrogens but differ from them in other respects. They are in common use in therapeutics because they are active by mouth and can be prepared at low cost.

The action of estrogens. The main effect of estrogens is stimulation of the development of the genital organs and mammary glands, an effect that is observed in intact and in castrated animals. In hens they stimulate the development of female plumage and inhibit the growth of the spurs.

In women and female monkeys, estrogens provoke proliferation of the endometrium, with an increase in the length of the uterine glands but with very little secretion; there are also vasodilatation and edema. If estrogen administration is discontinued, uterine hemorrhage occurs. This effect is seen in the infantile uterus and in women who have been castrated or have entered the menopause. In monkeys there is also reddening and swelling of the sexual skin around the vulva and anus, and water is retained. The tonus of the uterus increases, and it becomes more sensitive to the oxytocic effect of neurohypophyseal extracts.

The vaginal epithelium increases in thickness, it desquamates cornified cells, and its secretions increase. The motility of the fallopian tube increases, the cilia are more active, and the secretory activity of the mucosa is greater. The ducts of the mammary gland proliferate in all species, but there is proliferation of the alveoli only in some, *e.g.*, in the guinea pig, in which estrogen provokes the secretion of milk.[1]

Estrogens in moderate doses stimulate the secretion of the luteinizing, luteotrophic (prolactin), and adrenotrophic hormones of the anterior hypophysis; in large doses they have an inhibitory effect. In adequate doses they prevent the appearance of or control the cytologic changes in the anterior hypophysis produced by castration. Large doses given repeatedly cause an increase in the size of the anterior hypophysis, accompanied by deficiency in all the functions of the gland.

In the male, estrogens cause testicular atrophy due to inhibition of the hypophysis. They also provoke dilatation and keratinization of the prostatic utricle in rodents and monkeys. The prostate is considerably enlarged and causes urinary retention in the bladder and hydronephrosis. In man, estrogens have been used in the treatment of cancer of the prostate, in order to inhibit the growth of the epithelial cancerous cells (see Chap. 60).

Estrogens produce metabolic effects such as

[1] Ethylnilestradiol and the so-called "artificial estrogens" are more resistant to the inactivating action of the liver than natural estrogens and are therefore effective by mouth.

[1] Diethylstilbestrol stimulates the development of the mammary glands in cows and goats, even after castration.

(*a*) retention of Na, Cl, and water; (*b*) decrease in blood phosphate; (*c*) diminished excretion of Ca and P in some cases in which it is increased, and establishment of positive Ca and P balances.

Estrogenic activity is usually expressed in international units. An international unit is the specific activity of 0.1 μg of estrone (this corresponds to approximately ⅓ rat unit). The international unit of estradiol benzoate is the specific activity of 0.1 μg of a standard preparation of the substance. The activity of estrogens is measured by injecting the substance into castrated rats or mice and determining the minimum amount necessary to provoke a vaginal cycle or an increase in uterine weight.

FUNCTIONS OF THE CORPUS LUTEUM

The corpus luteum is formed and functions at each cycle, but its most important activity takes place during pregnancy and is prolonged during lactation. The main functions of the corpus luteum are the following:

1. It inhibits ovulation and the sexual cycle, which do not occur if there is an active corpus luteum.
2. It diminishes the tonus of the uterus and decreases its sensitiveness to oxytocic substances.
3. It stimulates proliferation of the endometrium and its glandular secretion, and prepares it for the implantation of the fertilized egg (progestational activity).
4. It maintains the nutrition of the embryo and fetus.
5. It stimulates complete development of the mammary glands.

Progestational and gestational changes in the endometrium. Toward the middle of the cycle, after the ovum has been expelled, the follicle is converted into the corpus luteum, which secretes a hormone called progesterone. The uterine glands, which have grown during the proliferative stage of the cycle, owing to the action of estrogens, are stimulated by progesterone to a maximum development, and they enter into a secretory phase. This premenstrual secretory activity is a preparation of the uterine mucosa for the nidation and nourishment of the fertilized egg. If the ovum is not fertilized it is not implanted in the uterus and the corpus luteum retrogresses. Involution of the corpus luteum is the cause of the menstrual breakdown of the endometrium and the commencement of the maturation of a new follicle (or set of follicles) and of another cycle.

In species in which ovulation is followed by pseudopregnancy, the progestational changes in the endometrium are more marked. Thus Ancel and Bouin observed endometrial changes in the female rabbit after copulation with a vasectomized male. The ovum could not, of course, be fertilized in these conditions, but a corpus luteum was formed. The endometrium proliferated (Fig. 280), and the mammary glands were considerably enlarged, due to growth of the ducts and alveoli. If the corpora lutea are removed this development ceases, and it may be prevented if the extirpation is performed sufficiently early.

If, during the luteal phase of the cycle, a thread is passed through the uterus so as to include the endometrium, or simply an incision is made in the endometrium, a tumoral mass grows in the damaged area. This tumor has a structure similar to the decidual membrane of the pregnant uterus, and it is known as a deciduoma. This effect does not occur unless there is an active corpus luteum, *i.e.*, it is not observed during the proliferative stage or after removal of the corpus luteum (Leo Loeb). A deciduomatal response can be obtained from the uterus of a castrated animal by the injection of progesterone after having provoked the development of the endometrium by treating it with estrogens.

The effect of progesterone on the endometrium and mammary gland is considerably enhanced by giving adequate doses of estrogen before and during treatment with progesterone. Larger doses of estrogen inhibit the activity of progesterone.

The continuance of pregnancy. Toward the end of the last century, Born remarked that animals in which a placenta was not formed did not have corpora lutea. Fraenkel showed that extirpation of the corpora lutea prevented nidation or provoked abortion in the rabbit. This fact has been observed in many other species (rat, dog, etc.) in which the corpus luteum persists throughout pregnancy. In other animals (*e.g.*, guinea pigs and monkeys) it is not indispensable after the first stages of pregnancy, and normally it is seen to retrogress before the end of pregnancy. In women the corpus luteum can be removed after the fourth or fifth month of preg-

nancy without provoking abortion. There is some proof that at this stage the placenta begins to produce luteal hormone.

Progesterone treatment carries pregnancy to term, preventing abortion, after castration in pregnant rabbits (Corner). In the absence of progesterone the uterine tone increases and by compression kills the fetus. Extrauterine pregnancy can proceed normally after castration in the rabbit.[1]

Pregnancy can be prolonged by the injection of progesterone or of pregnant women's urine, which contains luteinizing hormone and causes the development of new corpora lutea (Snyder). An abundant secretion of estrogen seems to be necessary to maintain the development of the corpus luteum during pregnancy and to cooperate with its action on the uterus.

Development of the mammary gland. After the development of the mammary gland has been stimulated by estrogens, the hormone of the corpus luteum completes this development by provoking the growth of the acini and preparing the gland for the secretion of milk.

Action on the sexual cycles. An active corpus luteum, or the injection of progesterone, prevents ovulation. At each cycle ovulation is retarded by the action of the corpus luteum; if it is removed, ovulation occurs earlier. During pregnancy the corpus luteum prevents ovulation and the possibility of another fertilization (superfetation) when the uterus is already occupied. This effect is produced by the inhibition of hypophyseal gonadotrophin secretion.

Action on the uterine muscle. The secretion of the corpus luteum stimulates the growth of the uterine muscle during pregnancy (trophic effect). It also decreases the tonus of the uterus and renders it less sensitive to the oxytocic effect of posterior hypophyseal extract (Knaus). Adrenaline relaxes the virgin uterus and provokes contraction of the gravid uterus in several

a

b

c

FIG. 280. Sections of rabbit's uterus. *a*, castrate; *b*, castrate treated with 12 IU of estradiol; *c*, castrate treated with estradiol as in *b*, followed by 1.5 mg. progesterone. (*Courtesy of Professor R. Sammartino.*)

[1] COURRIER, R., and R. M. A. COLONGE, *Compt. rend. Acad. sc.*, Paris, **230**, 1438, 1950.

species. This is due to the secretion of the corpus luteum, because progesterone treatment in a virgin guinea pig will cause the uterus to respond to adrenaline with contraction.

Action on the pubic symphysis. In certain species, *e.g.*, the guinea pig, the ligaments of the pubic symphysis and sacroiliac joints soften and relax during pregnancy. Three hormones can provoke this effect: (*a*) estrogen, when treatment is prolonged; (*b*) progesterone, after treatment with estrogen; (*c*) relaxin, extracted from the ovary, corpus luteum, and uterus, which has a much more rapid and intense effect than other hormones (Hisaw, 1926).

The hormone of the corpus luteum. Several active substances have been obtained from the corpus luteum.[1] The most active of these is progesterone (the names "progestin" and "luteosterone" are not now commonly used). It is a steroid related to the estrogenic, androgenic, and corticoadrenal hormones.[2]

[1] Mainly by Corner and Allen, 1929, Allen and Wintersteiner, 1934, Butenandt and Westphal, 1934, and Slotta, Ruschig, and Fels, 1934.

[2] Testosterone, desoxycorticosterone, and other steroids have a slight progestational activity. Ethylniltestosterone (pregneninolone) has a marked progestational effect, and it is active by mouth (Inhoffen and Hohlweg, 1938).

Progesterone

Pregnandiol

Pregneninolone

Progesterone has been synthesized by Butenandt from stigmasterol (a plant steroid found in soya beans) and from pregnandiol.

Pregnandiol is a waste product of progesterone which is excreted in the urine (Marrian) combined with glycuronic acid, forming sodium pregnandiol glycuronidate. This substance appears in the urine of normal women 12 days before the menses, *i.e.*, 1 or 2 days after ovulation, when the corpus luteum begins to secrete. Urinary excretion of this substance increases and reaches a maximum in about a week and then ceases suddenly one or two days before menstruation. During pregnancy large quantities of pregnandiol are eliminated in the urine, maximum amounts are found toward the eighth or ninth month. Ovariectomy does not suppress urinary elimination of pregnandiol in pregnancy, because the placenta continues to produce it. Women excrete in the urine as pregnandiol 10 to 20 per cent of progesterone injected into them. During pregnancy this rises to 30 to 35 per cent. The uterus is not indispensable for the conversion of progesterone to pregnandiol, but during the menopause it can contribute to this process.[1]

[1] SOMMERVILLE, I. F., and G. F. MARRIAN, *Biochem. J.*, **46,** 290, 1950.

The activity of progesterone and of extracts of the corpus luteum is expressed in international units; 1 IU is the specific activity of 1 mg. of standard progesterone. Activity is measured by the effect on the endometrium of adult rabbits, which have been submitted to sterile copulation (with a vasectomized male) and castrated immediately afterward; progestational proliferation is observed (Corner). Immature virgin rabbits can also be used, but the endometrium must first be sensitized by the injection of estrogens (Clauberg).

Progesterone has the following effects: (*a*) it causes nidation of the fertilized egg, and it prevents abortion after castration in rabbits; (*b*) it prevents abortion after hypophysectomy; (*c*) it produces progestational development of the endometrium, previously sensitized by estrogens, and pseudopregnancy in castrated animals; (*d*) it diminishes the tonus of the uterus and sensitiveness to the oxytocic effect of posterior hypophyseal extracts; (*e*) it reverses the response of the virgin uterus to adrenaline, causing contraction instead of relaxation; (*f*) it prolongs or suppresses the sexual cycle; (*g*) in certain species it prolongs pregnancy when given in large doses.

Progesterone has no effect on the endometrium unless the latter has been previously sensitized by estrogens.

Disturbances in ovarian functions. Ovarian insufficiency, sexual infantilism, and menstrual disorders have already been mentioned. Certain tumors of the granulosa have feminizing activity and provoke precocious puberty in girls or cause development of the uterus and the reappearance of the menstrual cycle in women after the menopause. Thecomas (tumors made up of theca cells) can also have a feminizing effect. Exceptionally thecomas and thecosis, *i.e.*, diffuse hyperplasia of the theca cells (Fraenkel, 1941), have a masculinizing effect, similar to that of arrhenoblastomas (see Chap. 57).

Therapeutic application of female hormones. Estrogenic hormones are given to obtain the development of the genital organs and mammary glands in cases of sexual infantilism. Large and repeated doses may inhibit the hypophysis. Estrogens are also used in cases of infantile vaginitis due to gonorrhea, in order to provoke thickening of the vaginal epithelium and an acid secretion, thus creating unfavorable conditions for the development of the germs. They are also used in cases of senile vulvar pruritis in order to modify the mucosa and secretions of the vulva. Estrogens have been used in the treatment of atrophic rhinitis. The main use of estrogens is the treatment of the symptoms of natural or artificial menopause (castration, x-rays, etc.) (see page 673). Examination of vaginal smears and biopsies of endometrium, at the appropriate times, after treatment show the changes characteristic of the various phases of the cycle. Estrogens are used to inhibit growth of cancerous cells in cases of cancer of the prostate.

Progesterone is used in the preventive treatment of abortion and to produce progestational development of the endometrium.

In cases of amenorrhea if the endometrium has undergone estrogenic development, treatment with progesterone (5 mg. daily for 5 to 6 days) will provoke menstruation within 4 days after the end of the treatment. If the endometrium is atrophic, gonadotrophins are given in order to stimulate the ovaries.

Androgens have been used in the treatment of hypermenorrhea and painful mastitis, but they should be used with caution as they have a masculinizing effect (growth of beard and body hair, acne, development of the clitoris, etc.).

Cancerigenic effects of estrogens. Estrogens are growth factors, therefore they can provoke the development of tumors in tissues that are sensitive to their morphogenic effect if hereditary factors and acquired conditions necessary for tumoral growth are also present.

Certain strains of mice have a high incidence of breast tumors. Castration prevents the development of these tumors (Loeb). In other especially bred strains, castration provokes the development of estrogenic corticoadrenal tumors, and in some cases tumors in the mammary gland. Certain strains have a high incidence of mammary tumors in the females, but not in males; tumors can be developed in the males of these strains by treatment with estrogens (Lacassagne). Large doses of estrogens given continuously for long periods to mice and guinea pigs cause the appearance of cancer in the uterus (Riesco). Prolonged administration of small doses of estrogens provokes the formation of fibrous tumors and subperitoneal fibromyomas in guinea pigs (Lipschütz *et al.*). Reports have been published of cases in which prolonged treatment with large doses of estrogens has been considered as an accessory factor in the development of tumors of the uterus and breast. On the other hand, it is worthy of note that some cancerigenic substances have estrogenic activity.

BIBLIOGRAPHY

ALLENDE, I. C. DE, and O. ORÍAS, "Cytology of the Human Vagina," Hoeber, New York, 1950.

BURROWS, H., and E. S. HORNING, "Oestrogens and Neoplasia," Blackwell, Oxford, 1952.

CALATRONI, C. I., V. RUIZ, and G. DI PAOLA, "Endocrinología Sexual Femenina," El Ateneo, Buenos Aires, 1947.

COURRIER, R., Interaction between Estrogens and Progesterone, *Vitamins & Hormones*, **8**, 179, 1950.

FARRIS, E. J., "Human Fertility," Author's Press, Palisades Park, N. J., 1950.

GARDNER, W. U., "Endocrinology of Neoplastic Diseases," Oxford, New York, 1947.

GARDNER, W. U., Studies on Steroid Hormones in Experimental Carcinogenesis, *Recent Progress in Hormone Res.*, **1**, 217, 1947.

HAMBLEN, E. C., "Endocrinology of Woman," Charles C Thomas, Springfield, Ill., 1945.

HEARD, R. S. H., The Metabolism of Estrogens, *Recent Progress in Hormone Res.*, **4**, 25, 1949.

HISAW, F. L., and M. X. ZARROW, The Physiology of Relaxin, *Vitamins & Hormones*, **8**, 151, 1950.

LIPSCHUTZ, A., "Steroid Hormones and Tumors," Williams & Wilkins, Baltimore, 1950.

MARAÑON, G., La Edad Crítica, 2d ed., Ruiz, Madrid, 1925

CHAPTER 60

The Endocrine Function of the Testes

The TESTES HAVE two functions: they produce the male gametes, *i.e.*, spermatozoa, and they secrete a hormone that develops and maintains the genital and extragenital male secondary sexual characters. Male genital secondary sexual characters or organs are (*a*) the vectors of the gamete, *i.e.*, vas deferens and penis; (*b*) glands such as the prostate, seminal vesicles, and Cowper's and urethral glands, which add their secretions to the sperm. Extragenital male secondary sexual characters are all those morphologic, functional, and psychic traits which differentiate the male from the female. The most typical psychic male character is the libido, *i.e.*, the sexual impulse toward the female. The relation between the testes and virility and the consequences of castration have been known for many centuries.

Physiologic histology. The testes are made up of seminiferous tubules, in which the spermatozoa are formed, and of interstitial epithelial cells (Leydig cells), which secrete the male hormone (Fig. 281).

The seminiferous tubules are formed by several layers of cells surrounded by a thin connective-tissue membrane. The cells of the deeper layers (spermatogonia) divide and produce the more superficially placed spermatocytes of the first order. These on dividing produce spermatocytes of the second order, which by division give rise to spermatids. The latter do not divide but are converted into spermatozoa. Between the spermatogonia there are long columnar cells which emerge into the lumen of the seminiferous tubules. These are the Sertoli cells, on which the spermatozoa are inserted until they are ready to be released and ejaculated. Spermatogenesis (*i.e.*, the formation of spermatozoa) begins at puberty; it is provoked and maintained by the gametogenic (follicle-stimulating) hormone of the anterior hypophysis.

The interstitial, or Leydig, cells originate in the mesoderm. There are large numbers of these cells in the fourth month of fetal life. At birth they are less numerous, and they continue to decrease throughout childhood. At puberty they increase, and there is a more or less constant number during adult life, until they diminish in old age. These cells are typical epithelial cells with mitochondria, granules, rods, and pseudocrystals in the protoplasm and with a well-developed nucleus. They are situated around the blood vessels. In mammals development of the Leydig cells is stimulated by the interstitial-cell-stimulating (luteinizing) hormone of the anterior hypophysis.

ENDOCRINE REGULATION OF TESTICULAR FUNCTIONS

In man, the rat, and other animals the testes are in permanent activity and spermatogenesis is continuous. In other species there are seasonal changes; periods of sexual activity alternate with others of sexual rest, during which spermatogenesis ceases and the sexual characters retrogress. In man spermatogenesis diminishes gradually between the ages of fifty and seventy and may cease completely, but spermatogenetic activity has been reported even in extreme old age (ninety years).

The internal secretion of the testes is produced by the interstitial cells (Ancel and Bouin, 1903). If the vas deferens is tied, the seminal tubes de-

generate and only the Sertoli cells remain; the interstitial cells, on the contrary, are hypertrophied (Fig. 281*b*). The secondary sexual characters are not disturbed by this operation. The same result can be obtained by irradiating the testicles with a moderate dose of x-rays; however, large doses can cause atrophy of the interstitial cells and signs of castration appear. In certain individuals the testes do not descend into the scrotum, but remain in the abdomen or inguinal canal (cryptorchism). The seminal tubes of the undescended testicles are more or less atrophied, but usually the interstitial cells are well developed and the secondary sexual characters, including the sexual impulse, are normal. In some of these cases, however, the secretion of testicular hormone is deficient and castration cells are found in the anterior hypophysis. In man the relationship between the state of the Leydig cells and the development of secondary sexual characters has been well established.

a *b*

FIG. 281. Sections of rabbit's testes. *a*, normal adult, active spermatogenesis, few interstitial cells; *b*, several months after vasectomy, no spermatogenesis, only Sertoli cells, hypertrophy of interstitial cells.

Androgenic hormones stimulate spermatogenesis, and in hypophysectomized animals they prevent atrophy of the seminal ducts and maintain the production of sperm. Certain nonandrogenic hormones produce this same effect. These hormones should not be used for this purpose in human therapeutics, because they inhibit gonadotrophin secretion by the anterior hypophysis (as is shown by the decrease in the urinary excretion of gonadotrophin) and may thus eventually depress spermatogenesis.[1]

The development and maintenance of sexual functions depend principally on the continuous action of gonadotrophins secreted by the pars distalis of the hypophysis. In mammals the gametogenic (follicle-stimulating) hormone provokes the development of the seminal tubes and spermatogenesis even in immature or hypophysectomized animals, but it does not stimulate the interstitial cells or the development of secondary sexual characters. The interstitial-cell-stimulating (ICSH) or luteinizing (LH) hormone stimulates the development and secretion of the Leydig cells and therefore the development of secondary sexual characters, but it has no direct effect on the seminal tubes of immature animals. In hypophysectomized animals

[1] HELLER, C. G., and W. O. MADDOCK, *Vitamins & Hormones*, 5, 393, 1947; HELLER, C. G., *et al.*, *J. Clin. Investigation*, 30, 648, 1951.

this hormone acts indirectly on spermatogenesis by stimulating the secretion of androgenic hormones by the Leydig cells.

Simultaneous administration of gametogenic (FSH) and interstitial-cell-stimulating or luteinizing (ICSH or LH) hormones provokes intense spermatogenesis and stimulates the Leydig cells not only in normal animals but also in hypophysectomized ones. Luteotrophin (prolactin) has no appreciable effect on the testes of mammals, but in birds it causes involution of the testicles.

Hypothalamus
Adrenals
Thyroid
Anterohypophysis
Cold
Light
Gametogenic gonadotrophin + (FSH)
− −
+ Interstitial cell-stimulating or luteinizing gonadotrophin (ICSH or LH)
Interstitial cells
Seminal tubes + Testicular hormone
Spermatozoa
+ 17-ketosteroids in urine

Secondary sexual characters
Morphogenic: genital (vectors of the gametes; glands of the genital tract); extragenital
Functional
Psychic: Virility; Libido

Hypophysectomy provokes testicular atrophy. Spermatogenesis ceases in the seminal tubes owing to the lack of FSH, and no more spermatozoa are ejaculated (azoospermia). The interstitial cells also atrophy, owing to the lack of ICSH, and the secondary sexual characters retrogress. An excess of ICSH or of chorionic gonadotrophin stimulates the testes in immature animals, causing them to grow and, when undescended, to migrate into the scrotum (Engle). The prostate and seminal vesicles increase in size. In the adult the testicle does not increase in size, having reached the fullest possible development, but the Leydig cells are stimulated, and the hypersecretion of androgenic hormones thus provoked causes the prostate, seminal vesicles, and other secondary sexual characters to grow.

The testes control and moderate the sexual function of the anterior hypophysis. Castration is followed by an increase in the gonadotrophin content of the hypophysis; there is also an increase in its secretion. The injection of androgens prevents hyperfunction of the anterior hypophysis and the appearance of the castration cells in the pars distalis. Prolonged treatment with large doses of androgens may inhibit gonadotrophin secretion sufficiently to provoke testicular or ovarian atrophy. Estrogens are much more efficacious than androgens as inhibitors of the anterior hypophysis.

A normal rate of thyroid secretion is necessary for the normal gonadotrophin secretion. Thyroidectomy is followed by a decrease in this secretion and subsequently by testicular atrophy. Severe hyperthyroidism also disturbs gonadotrophin secretion and indirectly the functions of the testes.

Several factors modify the structure and functions of the testes by acting on the hypophysis. Darkness and confinement cause testicular atrophy in some species. Intense illumination, on the contrary, provokes growth of the testes and secondary sexual characters in hibernating animals such as the ferret and in ducks.[1]

Deficiency of vitamin E and, sometimes, of other vitamins (A, B_1) provokes atrophy of the seminal tubes and Leydig cells, which becomes irreversible if the deficiency is severe. Prolonged undernourishment depresses testicular functions, and the same effect is produced by alcoholic excess. Testicular atrophy has been observed in animals fed on diets deficient in manganese. Hepatic lesions and the loss of bile through a fistula also cause testicular atrophy.

If the testes remain in the abdominal cavity without descending into the scrotum (cryptorchism) or if they are experimentally transplanted into the abdomen, the seminal tubes degenerate and spermatogenesis ceases. This has been attributed to the high temperature in the abdominal cavity. Sensitiveness of the testes to heat has been demonstrated by warming the scrotum, a procedure which causes degenerative changes in the seminal tubes. Patients with high and prolonged fever also show disturbances in spermatogenesis. These facts indicate that the

[1] In ducks irritation of the eyes has the same effect.

lodging of the testes in the scrotum outside the abdomen assures a lower and more suitable temperature for normal spermatogenesis than the higher intra-abdominal temperature.

The removal of one testicle and a large part of the other is not followed by signs of castration if a sufficient amount of interstitial cells remains. In animal experiments a residue of only 2 per cent of the whole testicular weight has been found to be sufficient. The Leydig cells of the remaining tissue undergo compensatory hypertrophy if the hypophysis functions normally, but not if there is hypophyseal insufficiency, unless gonadotrophins are injected.

Castration. The removal of the testes (castration) has been practiced since prehistoric times in order to tame animals or to improve their meat by softening and fattening. In man castration has been practiced by certain religious sects and to obtain eunuchs to serve as custodians of harems or male singers with treble voices. Castration is sometimes the consequence of trauma, and it has been performed in cases of testicular tumors and in order to retard the evolution of cancer of the prostate.[1] In some countries castration has also been legally performed with the object of preventing procreation by persons with hereditary diseases.

Prepuberal castration causes much more severe disturbances than castration in adults. Most of the secondary sexual characters do not develop, while in adults some of the sexual characters do not retrogress after castration and others undergo only an incomplete retrogression. Total castration produces a condition called eunuchism; the condition caused by testicular insufficiency is known as eunuchoidism. Persistence of prepuberal forms in adults who have been castrated in childhood is called sexual infantilism.

Persons castrated in adult life are of normal height. Prepuberal castrates[2] are usually tall; their height is definitely above the average (Fig. 282). They have long limbs and therefore a remarkably wide span. The width at the shoulders is less than normal. They have a narrow chest; on the other hand, the pelvis is wider than in females, and they have broad hips. The cranium is small, the face is broad, and the nose is usually not prominent. The abnormal height and length of limbs are due to persistence of activity in the epiphyseal cartilages beyond the age at which it usually has ceased.

The genital organs do not develop after prepuberal castration, and there is no libido. Castration in adults is seldom followed by a decrease in the size of the penis; the prostate and seminal vesicles are atrophied. The libido and erections persist in many subjects, especially if castration has taken place late in life. Copulation, with orgasm and ejaculation of fluid from the seminal vesicles and prostate but without spermatozoa, has been reported in men 10 to 25 years after castration.

Prepuberal castrates have no beard or mustache. Hair does not grow in the axilla, and the pubic hairs are scarce and have a feminine distribution. The hair on the head is usually abundant, fine, and straight. In adult castrates, the beard and mustache frequently continue to grow.

The skin of young castrates is at first fine and soft like the skin of children. Later it becomes sallow (an effect of carotenoids) and finely wrinkled and takes on a senile aspect. Seborrhea and acne do not occur in castrates. In many cases fat is deposited, especially on the hips and buttocks and in the mammary region.

The muscles are poorly developed and soft, and the subjects are easily fatigued. The larynx does not develop after prepuberal castration, and castrates retain all their life the treble voice of childhood, with something of falsetto. The voice does not change after castration in adults.

The urine of normal men contains androgenic substances (17 to 122 IU per day, with an average of 70 IU); in women the urinary elimination of androgens is approximately 70 per cent that of men.[1] There is also urinary elimination of 17-ketosteroids. The amount excreted by normal young men and women varies over a wide range and there is considerable overlapping in the values found in the two sexes. In men the majority of mean values reported lie between 12.5 and 16.7 mg. per 24 hr., with a lower limit of 6 mg. and an upper limit of about 25 mg. The range of mean values for women is approximately 7 to 12 mg. per 24 hr., with a lower

[1] Huggins, C., *Ann. Surg.*, **115**, 1192, 1942.

[2] Tandler, J., and S. Grosz, "Die biologischen Grundlagen der sekundären Geschlechtscharacter," Springer, Berlin, 1913; Pittard, E., "La Castration chez l'homme," Masson et Cie, Paris, 1934; Hamilton, J. B., *J. A. M A.*, **146**, 1903, 1941.

[1] Hamburger, C., *et al.*, *Acta pharmacol. et toxicol.*, **1**, 129, 1945; Heller and Maddock, *loc. cit.*; Schon, H. I., *Acta endocrinol.*, **8**, 149, 1951.

limit of 3 mg. and an upper limit of about 22 mg.[1] In childhood smaller amounts are excreted; at the age of puberty there is a sudden increase, and a maximum is reached between the twentieth and twenty-fifth year; in old age excretion of these substances diminishes. Approximately five-eighths of the 17-ketosteroids excreted is produced by the adrenal cortex and the rest in the gonads. Androgen and 17-ketosteroid excretion diminishes after castration or adrenalectomy, but these substances are found in the urine of castrated adrenalectomized monkeys; therefore there must be some other source besides the adrenal cortex and the gonads. In cases of corticoadrenal and testicular hyperfunction (tumors with endocrine activity), large quantities of these substances are excreted in the urine. Injection of gonadotrophins, adrenocorticotrophin, androgens, or corticoadrenal hormones is also followed by increased elimination of these substances (see Chap. 54).

Creatine is sometimes found in the urine of castrates, and if creatine is given to them, they retain less than normal subjects.

a b c

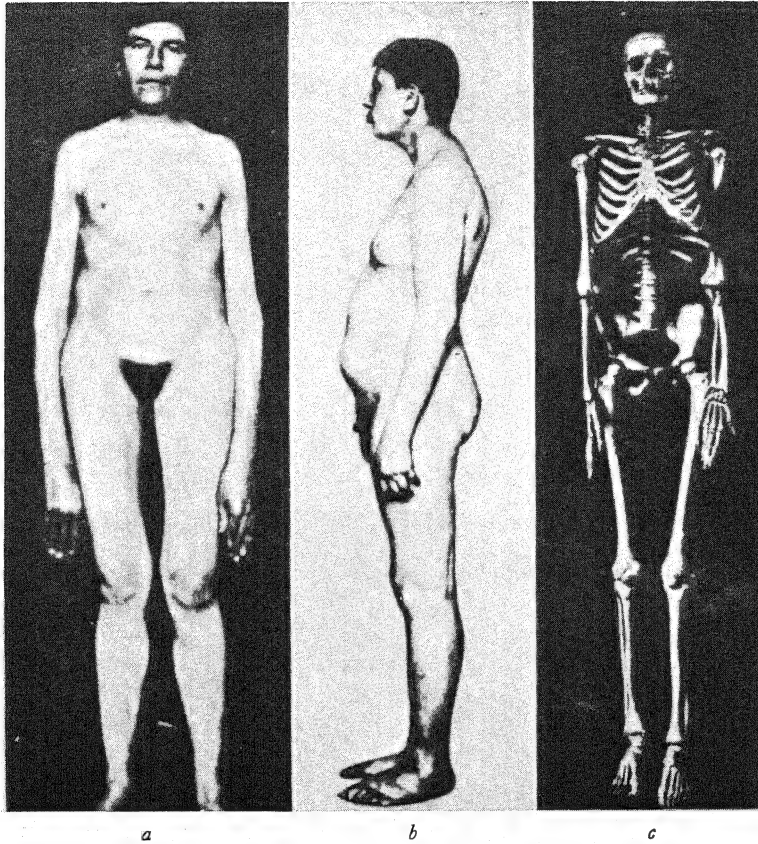



FIG. 282. Eunuchism. *a*, eunuch, twenty-four years old, castrated when five years old; height, 1.84 m.; long limbs, broad hips, no beard, feminine distribution of pubic hair. *b*, eunuchoid, twenty-six years old, no beard, fat on chest and hips. *c*, skeleton of eunuchoid, long limbs, epiphyseal cartilages not calcified.

The BMR is often subnormal in castrates, and the arterial blood pressure is low. In a few individuals vasomotor crises with sweating (hot flashes) have been observed. The thymus is usually enlarged.

The life span is apparently not shortened by castration, and in spite of the wrinkled aspect of the skin, there are no definite proofs that castrates suffer from premature senility.

Castration may provoke psychic disturbances owing to difficulties in social adjustment. Castrates are frequently timid and pacific, sober,

[1] MASON, H. L., and W. W. ENGSTROM, *Physiol. Rev.*, 30, 321, 1950.

honest, and hard-working, a fact that has been observed in the Skopzy, a religious sect in southern Russia and Rumania which practiced castration.[1] Intellectual development is not disturbed by castration, and some castrates are remarkably intelligent and of outstanding intellectual achievement.

Stimulation of the development of sexual characters. Deficiency in male secondary sexual characters can be corrected by (*a*) administration of interstitial-cell-stimulating (ICSH), or chorionic, gonadotrophin if there are Leydig cells capable of response (in castrates these substances have no effect); (*b*) administration of androgenic extracts or substances that have a direct peripheral action.

Symptoms of castration can be prevented from appearing, or they can be controlled, by (*a*) implantation of testicular tissue; (*b*) testicular grafts; (*c*) injection of testicular extracts; (*d*) injection of androgens. Subcutaneous or intramuscular implantation of testicular tissue or androgens produces effects during the time the active substances are being reabsorbed. Testicular grafts[2] do not take as easily as ovarian grafts. Successful grafting is followed by complete restitution of male characters in the castrated cock, rat, guinea pig, and other animals; in man usually only transitory effects have been observed. The method most frequently used in man is the administration of androgens, *e.g.*, testosterone and methyltestosterone.

Male hormones. Masculinizing or virilizing hormones are called androgens. The following are the principal androgens extracted from animal tissues or fluids: (*a*) *androsterone*, extracted from urine by Butenandt (1931–1932) and prepared synthetically by Ruzicka (1934); (*b*) *testosterone*, extracted from the testicle by David and Laqueur (1934), which is 7 to 10 times as active as androsterone; (*c*) *dehydroisoandrosterone*, found in the urine. The testes also contain Δ^5 pregneninolone, which is not androgenic but acts on spermatogenesis. Virilizing substances are also produced by (*a*) the adrenal cortex, from which *adrenosterone* and other androgenic steroids have been extracted (Fig. 252); (*b*) arrhenoblastomas of the ovary (Fig. 263). The nature of the active substances secreted by the ovarian tumors is not yet known. Several derivatives of natural androgens have been prepared by synthesis; one of the most frequently used in therapeutics is methyltestosterone, which is active by mouth. Natural androgens are rapidly inactivated mainly in the liver when given by mouth.

Testosterone

Androsterone

Dehydroisoandrosterone

Adrenosterone

[1] Koch, W., "Ueber die Russisch-Rumanische Kastraten Sekte der Skopzen," Carl Fischer, Jena, 1921.

[2] Berthold, 1849; Pézard, 1911; Steinach, 1912; Harms, 1914; Thorek, 1922; Lipschütz, 1924; etc.

Testosterone and other steroids (Δ_4 androstene-3,17-dione, 17-ketocholesterol) have been found in the blood of the testicular vein of the dog.[1]

The potency of androgens is expressed in international units; 1 IU is the specific activity of 0.1 mg. standard androsterone. This activity is measured by (*a*) the growth of the capon's comb; (*b*) prevention or control of atrophy of the prostate and seminal vesicles in rodents, estimated by the weight or the histologic aspect of these organs; (*c*) in guinea pigs, restoration of the capacity to ejaculate upon electrical stimulation of the nerve centers (in normal males, such stimulation provokes ejaculation of about 2 cc. of sperm which forms a clot; in castrates there is little or no secretion and it does not clot); (*d*) survival of spermatozoa in the testes of hypophysectomized rats or in the epididymis of castrated ones. There are other methods of estimating androgenic activity, but the most frequently used are (*a*) and (*b*).

Testicular hormones have sexual and general effects. The sexual effect consists in the development and maintenance of male genital and extragenital secondary sexual characters. In several species these hormones stimulate spermatogenesis even after hypophysectomy, but in man testosterone treatment diminishes spermatogenesis. The action of androgens secreted by the testes in this respect is not well known.

Male hormones produce important general metabolic effects. Nitrogen retention is increased and a positive nitrogen balance is established.[2] Phosphorus, potassium, sodium, sulfur, and other elements utilized for the formation of tissue are also retained. Muscular development is stimulated by male hormones, especially by testosterone propionate. The latter is used in cases in which nitrogen retention and the formation of cellular proteins are desired, or in which there is excessive nitrogen catabolism. Thus creatinuria is diminished or suppressed by testosterone.[3]

The sexual effect of androgens can be demonstrated (*a*) in the fetus, by stimulation of the development of male characters in male and female fetuses, in some cases provoking intersexuality (Chap. 57); (*b*) in infantile animals, in which precocious puberty is provoked; (*c*) in women, in whom male characters develop, such as growth of hair on face (beard) and body, lowering of the voice pitch, growth of the clitoris, suppression of menstruation,[1] decrease in milk secretion, acne, etc. (when the treatment is discontinued some of these changes retrogress); (*d*) in male castrates, in which the symptoms of testicular insufficiency are controlled; (*e*) in men, by general metabolic effects and stimulation of male characters.

Androgens given in adolescence at first increase growth in height, but ossification of the epiphyseal cartilages is completed prematurely and the subjects are not above normal height. Treatment at this age causes the penis, prostate, and seminal vesicles to develop more rapidly. If the testes were undescended they migrate into the scrotum, which grows and becomes pigmented. The beard, axillary, pubic, and body hair grow and have the aspect of the adult male. The skin thickens, seborrhea and acne occur, and frequently there is loss of hair on the head. The larynx grows and the register of the voice is lowered. The muscles develop and increase in strength and in resistance to fatigue. Libido, erection, and copulatory reflexes are established. The effects of castration on the anterior hypophysis are prevented or suppressed by androgens, which also virilize the adrenal cortex and diminish the size of the thymus. In the urine, androgens and 17-ketosteroids increase and gonadotrophins diminish if they had been previously increased by castration. (See Chap. 57 for the bisexual effects of androgens, and their antagonistic and synergic action with female hormones.)

Estrogens have been extracted from the testes of horses, pigs, and men. Estrone and estradiol are excreted in the urine of horses and men; they diminish after castration.

McCullagh, Martins, and Rocha postulated the existence of a testicular hormone that inhibited gonadotrophin secretion and prevented the appearance of the histologic changes in the anterior hypophysis following castration. This hypothetical hormone, called "inhibin," has never been identified. On the other hand, the prevention of castration changes in the hypophysis by testosterone has been demonstrated. There is experimental evidence which shows that the small amount of estrogen found in the testicle may be

[1] West, C. D., *et al.*, *Endocrinology*, **12**, 915, 1952.
[2] Kochakian, C. D., *Vitamins & Hormones*, **4**, 255, 1946.
[3] Methyltestosterone increases creatinuria.

[1] Hens cease laying eggs.

responsible for the prevention of the changes in the hypophysis following castration. Observations on pathologic conditions in man and animals suggest that the Sertoli cells may secrete estrogens. These cells are homologous to the granulosa cells of the ovary. Sertoli-cell tumors with feminizing activity have been found in dogs. Sertoli cells do not usually inhibit the hypophysis (Heller), but occasionally they may have an inhibitory effect (del Castillo *et al.*). The Leydig cells, however, seem to be the main source of estrogens in males (Maddock and Nelson).

Functional exploration of the testes. The functional condition of the testes can be determined by the following procedures: (*a*) examination of the development of male secondary sexual characters; (*b*) examination of the sperm, counting the spermatozoa and noting their motility; (*c*) biopsy and histological examination of small fragments of testis; (*d*) estimation of the gonadotrophin and 17-ketosteroids excreted in the urine.

TESTICULAR INSUFFICIENCY

Testicular insufficiency can be due to (*a*) primary disturbances or lesions in the testes; (*b*) disturbances in the anterior hypophysis which cause secondary testicular insufficiency. In both types of cases the age of onset is of importance, and the effects differ according to whether the disturbance began in childhood (prepuberal) or in the adult (postpuberal).

Primary testicular insufficiency. This may be either total or partial.

Total primary insufficiency is caused by traumatic or surgical castration (eunuchism), congenital absence of the testes (anorchism), the destruction of the testes by tumors or infections, etc. It is also observed in cases of so-called "male climacteric." In all these conditions the endocrine and spermatogenetic functions are diminished or absent. Urinary excretion of follicle-stimulating (gametogenic) gonadotrophin is increased; that of 17-ketosteroids is slightly diminished and does not increase after gonadotrophin injection.

Partial primary testicular insufficiency may involve either the seminal tubes or the interstitial cells.

Seminal insufficiency occurs in several forms:

1. Klinefelter's syndrome,[1] which begins at puberty. There is hyalinization of the seminiferous tubules; therefore spermatozoa are not formed (azoospermia). The Leydig cells are normal, and the secondary sexual characters may develop normally, but in some cases they are definitely subnormal. Gynecomastia, due mainly to proliferation of periductal connective tissue but also to growth of the ducts, occurs. The urinary excretion of gonadotrophin is increased as in castrates.
2. Seminal insufficiency in which spermatogenesis is not completed; there is therefore azoospermia. The secondary sexual characters are normal.
3. Cryptorchism, *i.e.*, undescended testes, in which the interstitial cells are well developed but there is no seminal tissue.
4. Sertoli-cell syndrome (del Castillo *et al.*),[1] in which the testes are slightly smaller than normal and there are no seminiferous cells, but the Sertoli cells are normal. The Leydig cells, the secondary sexual characters, and urinary excretion of 17-ketosteroids are normal. Urinary excretion of gonadotrophin may be normal, but it is usually increased (Nelson).

No cases have been reported of Leydig-cell atrophy with normal seminiferous tubules.

Secondary testicular insufficiency. Several forms due to hypophyseal insufficiency have been described; in all of them the Leydig cells and seminiferous tubules are atrophied: (*a*) total anterior hypophyseal insufficiency due to destruction of the anterior hypophysis by tumors such as craniopharyngiomas, by suprachiasmatic arachnoiditis, by trauma, etc.; (*b*) eunuchoidism, due to selective insufficiency of gonadotrophin secretion; (*c*) inhibition of the hypophysis by estrogens, in which case the testes are atrophied and the urinary excretion of gonadotrophins diminishes. In all these cases the urinary excretion of gonadotrophin and 17-ketosteroids is below normal. The administration of gonadotrophins increases excretion of 17-ketosteroids.

Testosterone treatment is given in cases of primary insufficiency due to castration. In man the most commonly used methods of administering this treatment are (*a*) subcutaneous injection of testosterone propionate (25 mg. two to five times per week); (*b*) oral administration of methyltestosterone (25 to 100 mg. per

[1] Klinefelter, H. F., E. C. Reifenstein, and F. Albright, *J. Clin. Endocrinol.*, **2**, 615, 1942.

[1] Del Castillo, E. B., A. Trabucco, and F. A. de la Balze, *J. Clin. Endocrinol.*, **7**, 493, 1947.

day). Other methods are used less frequently: (*a*) subcutaneous implantation of testosterone pellets (five pellets of 200 mg. produce effects that last several months); (*b*) sublingual absorption of testosterone (5 mg. dissolved in 0.2 cc. propylene glycol); (*c*) cutaneous absorption of testosterone applied in ointments.

Testosterone rapidly inhibits spermatogenesis. Prolonged treatment (25 mg. daily for 24 to 100 days) in normal men provoked first a decrease of spermatozoa, then azoospermia, with marked histologic lesions in the testes. There was no urinary excretion of gonadotrophins. When the treatment was discontinued, there was some improvement after 6 months and complete recovery in 17 months (Heller *et al.*). In some cases of hypogonadism, after the initial depression produced by testosterone, an increase in spermatogenesis above the primitive level has been observed (Heckel, Heller, *et al.*).

TESTICULAR HYPERFUNCTION

Injection of androgens, or of gonadotrophin if the testes are in condition to respond, provokes abnormally large development of some of the secondary sexual characters, *e.g.*, the prostate and seminal vesicles in rodents and the comb in cocks. These effects are not always sustained, and in some cases the organs retrogress after having been stimulated.

Several types of precocious puberty have been described in man.

1. Precocious puberty with premature hyperfunction of the Leydig cells.
2. Hyperplasia or tumors of Leydig cells (adenoma or carcinoma), with precocious puberty. The first case, described by Sacchi (1895), was that of a child of nine years with a tumor in the left testicle. The patient's growth was accelerated and he rapidly reached a height of 1.43 cm. The beard grew, the voice became low, and there were erections and ejaculations. After the tumor was removed the erections ceased, the beard was lost, and the voice regained the infantile treble. Eight similar cases have been reported.[1]
3. Constitutional syndrome of hypermasculinization, so called when the cause of the symptoms is unknown.
4. Tumors or lesions of the hypothalamus, hydrocephalus of the third ventricle, and tumors of the epiphysis and the neighboring structures.

[1] Sandblom, P., *Acta endocrinol.*, **1**, 107, 1948.

Rejuvenation. Methods for increasing sexual vigor, muscular strength, and intellectual alertness in old age are announced from time to time. After critical consideration of the results obtained, it is seen that they consist in transitory sexual stimulation, due largely to psychic causes (suggestion). The following methods of rejuvenation or sexual reactivation have been proposed: (*a*) injection of testicular extract (Brown-Séquard, 1889); (*b*) ligature of the vas deferens (Steinach, 1920), which produces atrophy of the seminiferous tubules and proliferation of the interstitial cells; (*c*) testicular grafts;[1] (*d*) testicular denervation (Dopler); (*e*) injection of blood taken from the testicular veins or of blood from young subjects; (*f*) implantation of monkey's testicle (Voronoff), which acts transitorily while active substances in the tissues are being reabsorbed.

BIBLIOGRAPHY

Allen, E., "Sex and Internal Secretion," 2d ed., Baillière, London, 1939.

Del Castillo, E. B., A. Trabucco, and F. A. de la Balze, Syndrome Produced by the Absence of Germinal Epithelium without Impairment of the Sertoli and Leydig Cells, *J. Clin. Endocrinol.*, **7**, 493, 1947.

Dorfman, R. J., Physiology of Androgens, and Biochemistry of Androgens, in "The Hormones," Academic Press, New York, 1948–1950.

Hamilton, J. B., The Role of Testicular Secretions as Indicated by the Effects of Castration in Man and by the Study of Pathological Conditions and the Short Life-span Associated with Maleness, *Recent Progress in Hormone Res.*, **3**, 257, 1948.

Heller, C. G., *et al.*, Effect of Testosterone upon the Human Testis, *J. Clin. Investigation*, **30**, 648, 1951.

Heller, C. G., and W. O. Nelson, Classification of Male Hypogonadism and Discussion on the Pathologic Physiology, Diagnosis and Treatment, *J. Clin. Endocrinol.*, **8**, 345, 1948.

Heller, C. G., and W. O. Nelson, The Testis-Pituitary Relationship in Man, *Recent Progress in Hormone Res.*, **3**, 229, 1948.

Hooker, C. W., The Biology of the Insterstitial Cells of the Testis, *Recent Progress in Hormone Res.*, **3**, 173, 1948.

Howard, R. P., R. C. Sniffer, F. A. Simmons, and F. Albright, Testicular Deficiency. A Clinical and Pathologic Study, *J. Clin. Endocrinol.*, **10**, 121, 1950.

Klinefelter, H. F., E. C. Reifenstein, and F. Albright, Syndrome Characterized by Gynecomastia, Asperma-

[1] Harms, 1914; Thorek, 1922; Voronoff, 1923. Critically reviewed by Moore, Velu, and others.

togenesis without a-Leydigism and Increased Secretion of Follicle-stimulating Hormone, *J. Clin. Endocrinol.*, **2**, 615, 1942.

MADDOCK, W. O., and W. O. NELSON, The Effects of Chorionic Gonadotrophin in Adult Man: Increased Estrogen and 17-Ketosteroid Excretion, Gynecomastia, Leydig Cell Stimulation and Seminiferous Tubule Damage, *J. Clin. Endocrinol.*, **12**, 985, 1952.

MASSON, H. L., and W. W. EDSTROM, The 17-Ketosteroids: Their Origin, Determination and Significance, *Physiol. Rev.*, **30**, 321, 1950.

NELSON, W. O., and C. H. HELLER, The Testis in Human Hypogonadism, *Recent Progress in Hormone Res.*, **3**, 197, 1948.

NEW YORK ACADEMY OF SCIENCE, Biology of the Testes, *Ann. New York Acad. Sc.*, **55**, 543–742, 1952.

CHAPTER 61

Fertilization. Pregnancy. The Physiology of the Fetus. Parturition

FERTILIZATION

Fertilization consists in the union of the spermatozoon and ovum. The optimum period for fertilization occurs in women around the middle of the menstrual cycle, when ovulation takes place. Normally sperm is deposited in the vagina by the erect penis, but artificial insemination, depositing the sperm in the uterine cavity by means of a catheter, has been performed successfully.

Erection. In mammals fertilization occurs within the body of the female (internal fertilization). Copulation is preceded by erection. The penis lengthens and hardens and from a flaccid, pendant structure is converted to a rigid and erect organ. These changes are due to the accumulation of blood in the erectile tissues of the penis, *i.e.*, the two corpora cavernosa penis and the corpus cavernosum urethrae, which expands into the glans. If an erect penis is tied at the base it remains erect until the ligature is removed; when the blood can flow out the penis becomes flaccid. Distention of the erectile tissue is due to dilatation of the afferent arteries; it is not produced by ligature of the veins of the penis. Vasodilatation during erection is demonstrated as follows: (*a*) if the corpora cavernosa are cut in the dog and the nervus erigens (formed by the first two or three sacral rami) is stimulated, blood flows from the wound in large quantities while the nerve is stimulated, and the flow diminishes considerably when stimulation ceases; (*b*) venous blood flow increases 10 to 15 times when stimulation of the nervus erigens is discontinued, *i.e.*, when erection retrogresses; (*c*) the blood pressure falls in the arteries and rises in the veins of the penis during erection; (*d*) plethysmography of the penis shows dilatation.

Erection is therefore primarily due to vasodilatation reflexly stimulated by cholinergic fibers of the parasympathetic. The most frequent stimuli that cause erection are cutaneous stimuli, especially those applied to the genital region and, more particularly, on the mucosa of the glans. Friction of the penis produces first erection and then ejaculation. Other cutaneous stimuli and visual and olfactory stimuli can also produce erection if they have been associated with the sexual impulse; psychic representation of appropriate stimuli is sufficient to provoke erection. Certain stimuli inhibit erection. A frequent cause of this inhibition is emotional stress, and many cases of impotence in young persons are exclusively due to psychic inhibition.

The centripetal path of reflex erection initiated in the penis goes along the dorsal nerve of the penis, a branch of the internal pudendal nerve. Section of this nerve prevents erection in some animals, *e.g.*, the horse, even when there is intense sexual excitement. The centrifugal path lies in the nervus erigens; the fibers afterward travel in the internal pudendal nerve. Stimulation of these fibers provokes erection, but not in castrates. In man fibers from the second to fourth sacral nerves conduct impulses which provoke erection. Extirpation of the thoracic sympathetic produces disturbances in erection in 57 per cent of the cases (Whitelaw *et al.*).

Ejaculation. Spermatozoa are being formed continuously. They are carried along the genital ducts by the cilia of the epithelia lining the

epididymis and the vas deferens and by their own motility. Thus they advance 5 to 7 m. and are deposited in the ampulla of the vas deferens. The seminal vesicles are not reservoirs for spermatozoa, although in some species spermatozoa may be found in them; they are glands which, together with the prostate, Cowper's glands, and the epithelia of the genital ducts, secrete a fluid that forms a substantial part of the ejaculate.

In the process of ejaculation, contractions of the vas deferens empty the sperm through the ejaculatory ducts into the urethra by two orifices situated on each side of the verumontanum. Simultaneously the seminal vesicles, Cowper's glands, and the prostate pour out their secretions into the urethra. The verumontanum becomes turgid and the sphincter vesicae contracts, preventing the flow of sperm into the bladder. Seminal fluid is not immediately expelled from the urethra, but accumulates in the membranous segment of the urethra between the verumontanum and the sphincter urethrae, which is contracted, damming up the fluid coming into the urethra under pressure, due to the contraction of the smooth-muscle fibers of the vas deferens, seminal vesicles, and prostate. The sphincter urethrae suddenly relaxes and the bulbocavernosus and ischiocavernosus muscles and the levator ani contract rhythmically, alternating with the contraction and relaxation of the sphincter urethrae. Thus seminal fluid is ejected intermittently into the vagina.

Ejaculation is a reflex phenomenon. The main impulses arise in the glans and in the urethral mucosa. The ejaculatory center is situated in the spinal cord at the level of the fourth and fifth lumbar segments in the rabbit, slightly cephalic to the center for erection. Efferent fibers leave the upper lumbar segments and form part of the adrenergic lumbar sympathetic; these fibers provoke contraction of the vas deferens and the seminal vesicles. Other fibers, arising in the third and fourth sacral segments, travel along the internal pudendal nerves to the ischiocavernosus and bulbocavernosus muscles and the sphincter urethrae. Ejaculation does not take place if the parasympathetic (sacral) fibers are destroyed. Emission of sperm is suppressed by destruction of sympathetic (lumbar) fibers; sometimes the sperm is sent into the bladder instead of being evacuated through the meatus. Removal of the lumbar sympathetic below the twelfth dorsal segment produces disturbances in ejaculation (Whitelaw).

Erection normally precedes ejaculation, but emission of seminal fluid occasionally takes place without erection. In the guinea pig ejaculation can be provoked by stimulation of the hypogastric plexus or the inferior mesenteric ganglion, which causes contraction of the vas deferens and seminal vesicles; also by electrical stimulation of the brain. Asphyxia provokes ejaculation in several species, particularly in rodents.

Ejaculation is accompanied by sensations and motor phenomena which constitute the orgasm. It is followed by muscular relaxation; the erectile organs become flaccid, and often there is a sensation of fatigue.

The amount of seminal fluid emitted at each ejaculation varies from 2 to 5 (average 3.2 cc.), with 40 to 300 million spermatozoa per cubic centimeter (average 90 million).[1] The quantity of seminal fluid ejaculated is closely related to frequency of copulation, diminishing as frequency increases.

Normal forms are found in 60 to 100 per cent of spermatozoa (average 82 per cent in 100 fertile men). Motility is seen in 40 to 90 per cent (average 61 per cent). Minimum conditions for fertility are: 2 cc. of sperm with 20,000,000 spermatozoa per cubic centimeter; 40 per cent with motility 3 of the Hotchkiss-MacLeod scale; normal forms in 60 per cent (MacLeod and Gold). Motility is the characteristic most closely related to fertility.

Chemical analysis of sperm[2] has shown it contains relatively large amounts of fructose (200 to 800 mg. per 100 cc.); choline, phosphocholine, citric acid, and acid phosphatase have also been found. Fructose and choline come from the seminal vesicles; acid phosphatase and citric acid come from the prostate. These substances diminish after castration, and increase during treatment with androgens.

Ejaculated seminal fluid clots in many animals, especially in rodents. The semen of guinea pigs and rats clots and forms the vaginal plug. This plug is supposed to prevent the escape of sperm from the

[1] MacLeod, J., *Ann. Rev. Physiol.*, **5**, 399, 1943; "Proceedings of the Conference on Diagnosis of Sterility," Charles C Thomas, Springfield, Ill., 1946.

[2] Mann, T., and C. Lutvak-Mann, *Physiol. Rev.*, **31**, 27, 1951; Lunquist, F., *Acta physiol. Scandinav.*, **19**, Suppl. 66, 1949.

vagina, or to stimulate the cervix and retard the involution of the corpora lutea (pseudopregnancy). It is formed by clotting of the fluid of the seminal vesicles brought about by the secretion of the prostate (guinea pig) or the coagulating glands (rat). A drop of this secretion can coagulate a large amount of the fluid of the seminal vesicles. Later the clot retracts and a clear serum separates. The coagulating ferment has been called vesiculase (Camus and Gley, 1896).

Accessory genital glands. The seminal vesicles add their fluid to the sperm in the process of ejaculation. Extirpation of these organs does not prevent normal copulation, but it diminishes fertility. Removal of the abdominal sympathetic nerve chains prevents evacuation of the seminal vesicles and prostate and in some species causes sterility.

The prostate also adds its secretion to the sperm.[1] After the age of forty it frequently increases in size, and it has been found hypertrophied in 30 to 50 per cent of men above the age of fifty-five. Hypertrophy involves not only the prostate but also other structures such as the periurethral glands. The enlarged prostate is usually an obstacle to the normal evacuation of the bladder and in some cases has to be removed in order to prevent urinary retention.

The prostate has been supposed to have endocrine functions essential for reproduction or to maintain the vitality of the sperm. This internal secretion has never been satisfactorily demonstrated. Moreover the prostate has been removed in animals and man without diminishing *potentia coeundi* (*i.e.*, copulatory capacity) or causing any disturbance in the sexual characters. *Potentia generandi* (*i.e.*, fertility) is sometimes impaired, probably owing to mechanical obstacles to the emission of sperm created by prostatectomy, but it is doubtful whether the absence of prostatic secretion plays a part in this. Prostatic fluid or extracts have several properties such as producing a fall in blood pressure when injected, or provoking contraction in the isolated intestine or uterus.[2] There is no proof that the substances that have these effects are secreted into the blood or have any importance in physiologic processes.

The development of the prostate is governed by testicular hormones. Prepuberal castration

[1] Huggins, C., *Physiol. Rev.*, **25**, 281, 1945; *Harvey Lect.*, **42**, 148, 1947.
[2] Euler, U. S. von, *Skandinav. Arch. f. Physiol.*, **81**, 65, 1935.

causes atrophy of the prostate. The hypertrophied prostate of older subjects also diminishes in size after castration. Testosterone treatment causes prostatic growth and hypertrophy, especially if administered before puberty. Cancer of the prostate is stimulated by androgens; on the other hand its development is retarded, and the tumor may decrease in size, after castration or injection of estrogens.[1] Castration and estrogen treatment have been applied successfully in order to alleviate pain and prolong life in cases of cancer of the prostate. Adrenalectomy has also been performed in some cases in order to suppress an additional source of androgens.

Pilocarpine stimulates and atropine inhibits the secretion of prostatic fluid. Stimulation of the hypogastric nerves provokes secretion of prostatic fluid and contraction of the muscle fibers of the prostate. Stimulation of the pelvic nerves contracts prostatic muscles but has little effect on the emission of prostatic fluid.[2] The secretory innervation of the prostate seems to be provided by sympathetic cholinergic fibers similar to those which stimulate the sweat glands. The prostate contains acid phosphatase, which increases in the blood of patients with cancer of the prostate.

Excitability and contractility of the smooth muscles of the vas deferens, seminal vesicles, and prostate are diminished by testosterone and increased by estrogens (Martins, Valle, and Porto).

Copulation in the female. The erect penis is introduced into the vagina, and there it ejaculates the seminal fluid. Sexual excitement during copulation causes erection of the corpora cavernosa of the clitoris and the bulbus vaginae, which are distended with blood. The clitoris does not rise in the process of erection; its extremity curves downward and is applied on the dorsal aspect of the penis. The bulbus vaginae is in part surrounded by the constrictor muscle of the vagina; rhythmic contractions of this muscle and turgescence of the bulbus vaginae tend to narrow the vulvar orifice of the vagina and close it on the penis. The bulbocavernous muscle and the levator ani also take part in producing this effect. Rhythmic movements of the pelvis frequently occur. In some animals the uterus and the os contract and relax

[1] Huggins, C., *Cancer Research*, **1**, 293, 1941; *Ann. Surg.* **115**, 1192, 1942.
[2] Farrel, J., *J. Urol.*, **39**, 171, 1938.

rhythmically during copulation. Orgasm is usually more prolonged in the female than in the male; it is accompanied by generalized motor activity and is followed by relaxation as in the male.

Erection of the clitoris does not always occur, or it may occur late in copulation. The female can be fertilized without having experienced erection or orgasm, as is demonstrated by the facts that an unconscious woman can be fertilized and that fertilization by artificial insemination can occur.

In the female, as in the male, the nervi erigentes are constituted by cholinergic parasympathetic fibers which arise in the first, second, and third sacral segments (in the dog).

The ducts of the glands of Bartholin open on each side of the vagina. During copulation these glands secrete a fluid that apparently serves to lubricate the penis and facilitate penetration.

PREGNANCY

Fertilization and nidation. Fertilization usually takes place in the fallopian tube. The head of one spermatozoon penetrates into the ovum, and there is fusion of the nuclei of the two gametes. The fertilized egg then divides, and the embryo is formed by the successive divisions of the egg. The ovum survives only 6 to 20 hr. after ovulation, and spermatozoa 30 to 48 hr. after ejaculation. The egg migrates down the fallopian tube and normally becomes implanted in the endometrium 6 or 7 days after fertilization.[1] The earliest human embryos that have been observed are one 4 days old, which had not yet been implanted, and one 7½ days old, already implanted (Rock and Hertig). In the process of nidation the embryo penetrates into the uterine mucosa (interstitial implantation). The trophoblast (the membrane covering the embryo) apparently releases a proteolytic enzyme that facilitates this process by destroying the adjacent mucosa.

Parthenogenetic division of the ovum has been provoked *in vitro* in the rabbit. If the ovum thus activated is placed in the uterus it can become implanted and develop into rabbits which are normally delivered (Pincus and Shapiro, 1940).

[1] Nidation occasionally occurs in the fallopian tubes, on the surface of the ovary, or in the abdominal cavity (ectopic pregnancy).

Fertilized eggs of one female have been transplanted into the uterus of another which had been prepared by hormonal treatment. The embryos developed to full term, in some cases even after several hours intervened between extraction and implantation of the egg. The fertilized ovum of man and the first stages of its development have been observed *in vitro* (Rock *et al.*, 1944).

It has been claimed that sperm must have a sufficient amount of hyaluronidase for the spermatozoa to penetrate into the ovum.

The duration of pregnancy. Pregnancy in women has a duration of approximately 280 days (40 weeks, *i.e.*, 9 months and 10 days) counting from the first day of the last menstruation.[1] Cases of normal babies born after a pregnancy of only 250 days or as long as 330 days have been reported. A pregnancy is considered abnormally prolonged, and the life of the fetus is in danger, when it lasts more than 300 days. In many species the duration of pregnancy is a multiple of the duration of the estral cycle.[2]

Changes in the uterus. The uterus grows considerably during pregnancy. The lumen of the virgin uterus has a capacity of 2 to 5 cc.; at full term this has increased to 5,000 or even 7,000 cc. The virgin uterus weighs 50 gm.; at the end of gestation it weighs approximately 1,000 gm. Uterine hypertrophy extends to all its parts and is especially marked in the myometrium. Each uterine muscle fiber increases to between 7 and 11 times its original length and 3 to 5 times its original transverse diameter. The number of uterine muscle fibers also increases (hyperplasia). Uterine hypertrophy is due to the action of estrogen[3] and to distention.[4]

Uterine contractility in the pregnant woman has been studied by registering pressure changes

[1] The average duration of pregnancy (in days) in different species is as follows: rat, 22; rabbit, 29 to 30; dog, 59 to 63; guinea pig, 60 to 66; cat, 63; sheep, 147 to 154; chimpanzee, 236; cow, 285; mare, 340; elephant, 660.

[2] Snyder, F. F., *Physiol. Rev.*, **180**, 578, 1938; Kenneth, J. H., "Gestation Periods," Oliver & Boyd, Edinburgh, 1943; Asdell, S. A., "Patterns of Mammalian Reproduction," Comstock Publishing Company, Inc., Ithaca, 1946.

[3] The stimulating action of estrogens is also seen during estrus in many animals and in female castrates treated with estrogens.

[4] Reynolds, S., *Physiol. Rev.*, **17**, 304, 1937; "Physiology of the Uterus," Hamilton, London, 1939.

in the amniotic fluid with a water manometer[1] (Fig. 283). During the first half of pregnancy there are continuous, small, rhythmic, regular, painless contractions which have been recorded from the second month. These contractions increase amniotic pressure by 1 to 7 cm. H_2O from a basal pressure of 4 to 11 cm. H_2O; their frequency is one to two per minute (Fig. 283*A*). Basal tonus has slow oscillations. Contractions of greater amplitude appear occasionally during the first months; they increase in frequency as pregnancy advances (Fig. 283*B*). During the last two weeks contractions increase in amplitude (Fig. 283*C*) and there is a gradual transition from the weak contractions of pregnancy to the strong contractions of labor.

The uterus becomes gradually more sensitive to the oxytocic effect of posterohypophyseal extracts. The increase in the excitability of the uterus toward the end of pregnancy can be demonstrated, not only in the uterus *in situ*, but also in the isolated surviving uterus.[2]

Three factors cause the increase in uterine excitability as pregnancy advances: (*a*) gradual increase in estrogens, which are uterine stimulants; (*b*) gradual decrease of progesterone, which inhibits the uterus; (*c*) distention of the uterus, which stimulates it. Adrenaline causes relaxation in the virgin uterus of certain animals, but on the contrary it provokes the contraction of the pregnant uterus. This inversion of the effect of adrenaline is due to the sensitizing effect of progesterone.

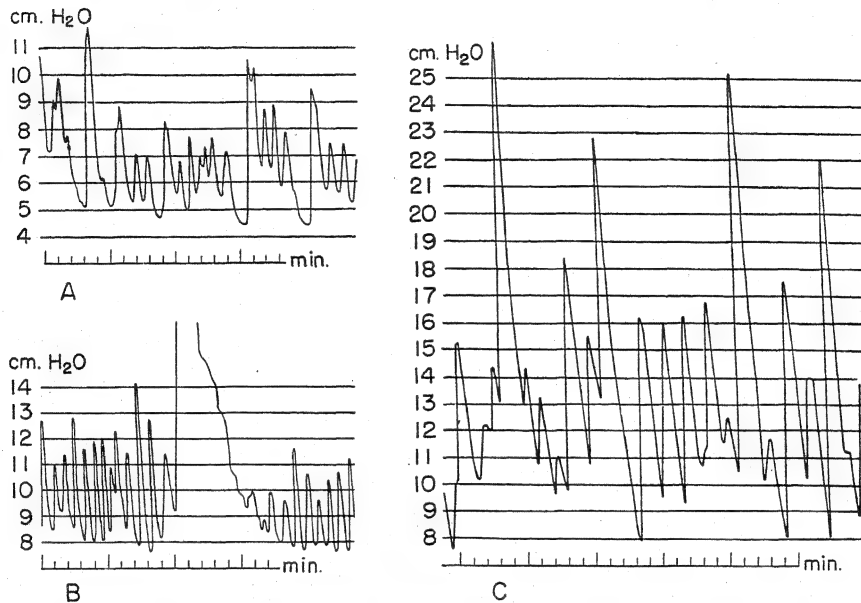



FIG. 283. Records of uterine contractions in pregnant women. *A*, 9 weeks; *B*, 7 months; *C*, 2 days before labor. (*Alvarez and Caldeyro-Barcia.*)

The uterine mucosa increases in thickness and vascularization during pregnancy. On the fifteenth day of pregnancy it is 0.5 cm. thick, and 0.7 cm. on the sixtieth day. The stroma cells are converted into large epithelioid cells, known as "decidual cells" because the uterine mucosa is shed after parturition and for this reason is called the *decidua*.

When the embryo penetrates into the endometrium a chamber lined by mucosa is formed. The mucosa that lines the uterine cavity outside the embryo is called the *parietal decidua* (*decidua vera*). The mucosa that covers the embryonic chamber and separates it from the lumen of the uterus is the *reflex decidua* (*decidua reflexa*). The embryo is separated from the uterine wall by the *basal decidua* (*decidua serotina* or *basalis*) (Fig. 284).

The parietal decidua is thick and well vascularized at the beginning of pregnancy, but

[1] ALVAREZ, H., and R. CALDEYRO-BARCIA, *Arch. urug. de ginecol. obst.*, **7**, 79, 1948; *Surg., Gynec. & Obst.*, **91**, 1, 1950.

[2] The human uterus is more sensitive to vasopressin than to oxytocin; the guinea pig uterus, on the contrary, is more sensitive to oxytocin (Scott, 1943).

from the fourth month on, it diminishes in thickness and at full term it is only 1 to 2 mm. thick. The reflex decidua also diminishes in thickness and by the sixth week only traces of it can be found; in the course of the fifth month its fusion with the parietal decidua is complete. The basal decidua takes part in the formation of the placenta.

The placenta. The placenta is the organ by means of which the fetus is nourished. It is formed of a maternal part (the basal decidua) and a fetal part (the chorionic villi). The fetal part of the placenta arises from the trophoblast, *i.e.*, the primitive membrane that surrounds the embryo. As the embryo develops, processes appear on the surface of the trophoblast which penetrate like fingers into the basal decidua. These processes are invaded by blood vessels and mesodermal tissue, and eventually they develop considerably, forming the chorionic villi. The villi grow into the vascular spaces of the basal decidua, and here again the cytolytic activity of the embryonic structures facilitates penetration into the maternal tissues. Fetal blood circulates through the villi, separated from the maternal blood by a double layer of epithelium—one layer of maternal, the other of fetal, origin. Later the two layers are fused into a single layer.[1] This epithelium is the barrier between the fetal and maternal circulations and the site of metabolic exchange between the mother and fetus. It is endowed with selective permeability.[2] Crystalloids pass through the membrane; colloids and small particles (bacteria, etc.) are retained. Oxygen, water, salts, foodstuffs, vitamins, and certain hormones pass into the fetus. CO_2 and waste products pass out from the fetus. Certain antibodies can also pass through the placenta into the fetus. The permeability of the placenta to radioactive sodium increases as pregnancy advances.[3]

The placenta is not only the site of exchange between the mother and fetus; it is also an endocrine gland, which produces hormones necessary for the continuance of pregnancy.

[1] The chorionic villi are lined by two layers; one layer is formed by the cells of Langhans, the other is a syncytium. The former gradually diminishes and in parts disappears completely.

[2] NEEDHAM, J., "Chemical Embryology," Macmillan, New York, 1931; SCHLOSSMANN, H., *Ergebn. d. Physiol.*, 34, 741, 1932.

[3] FLEXNER, L. B., and A. GELLHORN, *Am. J. Obst. & Gynec.*, 43, 965, 1942.

MATERNAL CHANGES IN PREGNANCY

In the course of pregnancy many important changes take place in the maternal organism in order to adjust it to the needs created by the development of the fetus.[1] The fallopian tubes, uterus, and vagina hypertrophy under the influence of estrogens and progesterone; the mammary gland and the nipple also develop. Pigmentation appears on the skin of the face, the abdomen, and the areola around the nipple. The abdominal viscera are displaced, and the diaphragm is pushed upward. The veins of the lower limbs and pelvis dilate and, in some cases, become varicose (hemorrhoids).

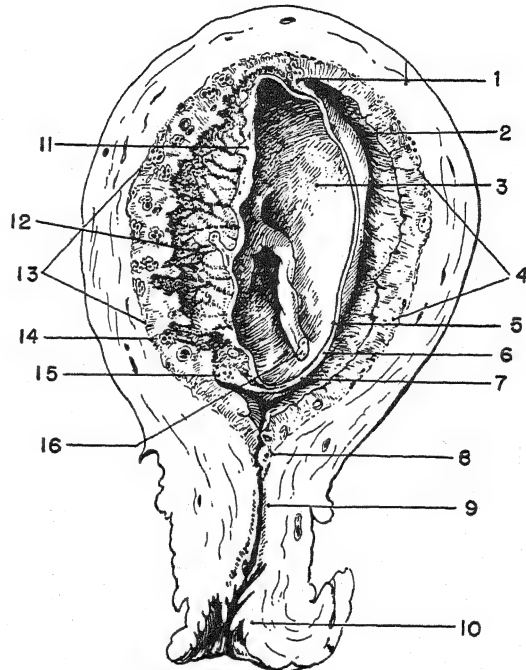



FIG. 284. Section of pregnant uterus with membranes of 12-week-old embryo. 1 and 15, site where the decidua is reflected on the embryo; 2, cavity of the uterus; 3, embryonic cavity; 4, parietal decidua (decidua vera); 5, amnion; 6, chorion levis; 7, decidua reflexa; 8, internal orifice of the cervical duct; 9, cervical duct; 10, external orifice of the cervical duct (os); 11, chorionic membrane; 12, chorion frondosus; 13, basal decidua; 14, dilated blood vessels of the basal decidua; 16, umbilical cord.

[1] HARDING, V. J., *Physiol. Rev.*, 5, 279, 1925; SEITZ, L., and R. T. JASCHKE, in Bethe's "Handbuch der normalen und pathologischen Physiologie," 14, 436 and 579, Springer, Berlin, 1926; VIGNES, H., in G. H. Roger and L. Binet, "Traité de physiologie normale et pathologique," vol. 11, Masson & Cie, Paris, 1934, p. 271.

An outstanding character of metabolism during pregnancy is the predominance of anabolic processes. Nitrogen, calcium, phosphorus, iron, water, and many other substances are retained by the maternal organism in order to provide the fetus with materials for building up its tissues. When considering the nutritional requirements of the human organism (Chap. 50), the increased needs of the pregnant and lactating woman were emphasized. Wherever there is the risk of nutritional deficiency, such as that of iodine in goitrous districts or that of vitamin D in northern industrial towns, care should be taken to provide sufficient amounts of the potentially deficient factor. It is necessary to bear in mind that intrauterine nutritive deficiency may produce irreversible unfavorable effects on the development of the fetus (*e.g.*, cretinism).

In the course of the last three months of pregnancy the maternal organism is submitted to considerable metabolic strain. During this period the fetus makes up 75 per cent of its protein, 93 per cent of its fat, 65 per cent of its Ca, 68 per cent of its P, 80 per cent of its iron, and 70 per cent of its weight at birth (Hugget, 1942).

The BMR of the mother increases toward the end of pregnancy. This increase is not due to an increase in maternal metabolic rate, but to the metabolism of the fetus. If the metabolisms of the mother and child are determined shortly after birth, it will be seen that the sum of both adds up to the figure found in the mother just before birth.

The alkali reserve diminishes in pregnancy. Blood and alveolar CO_2 tensions are below the normal, and there is marked hyperventilation. It is still unknown whether hyperventilation precedes or follows the decrease in the alkali reserve. This phenomenon has been called the "physiologic compensated acidosis of pregnancy," but the hydrogen ion concentration is always normal.

Liver glycogen falls more rapidly during fasting in pregnant animals than in nonpregnant ones. Blood fat is high, and total blood cholesterol (especially free cholesterol in plasma) rises, although in the erythrocytes it may fall. Creatinuria is often observed.

The minute output of the heart increases rapidly from the tenth week, reaching a maximum at the twenty-sixth to twenty-ninth week (30 per cent above the basal nonpregnant level); later it diminishes.[1] Extracellular fluids (blood plasma and tissue fluid) increase; therefore the circulating volume is also increased, and erythrocyte concentration diminished, especially toward the end of pregnancy. Neutrophil leukocytosis is often seen. The suspension stability of the red blood cells is considerably diminished; the erythrosedimentation rate is therefore increased (Fåhraeus). This is due to the increase in plasma globulins (see Chap. 3).

Toward the end of pregnancy, 35 to 60 per cent of women have slight glycosuria or lactosuria, which ceases soon after delivery.

Pregnant women retain water easily. Toward the end of pregnancy the urine should be periodically examined for the presence of albumin, the blood pressure should be determined, and the appearance of edema or a sudden increase in weight should be watched. These are premonitory symptoms of eclampsia, a convulsive condition with a serious prognosis, which can usually be avoided by an adequate diet and by restricting the intake of salt and water. Plasma clearance tests show that there is frequently some impairment of renal function during the last stages of pregnancy.

There is a considerable increase of histaminase in blood; apparently this enzyme is produced by the decidua. Oxytocinase also increases.

Hormonal changes in pregnancy. The anterior lobe of the hypophysis hypertrophies, and typical cells (pregnancy cells) appear. The gonadotrophin content of the gland diminishes considerably in pregnant women. The thyroid and the adrenal cortex increase in weight, and the thymus undergoes involution. The urine of pregnant women contains large quantities of estrogens, chorionic gonadotrophin, and pregnandiol.

These estrogens are mainly estriol and smaller amounts of estrone and estradiol bound to glycuronic acid. They play an important part in the processes of pregnancy by (*a*) stimulating the growth of the uterus and the rest of the genital tract; (*b*) sensitizing the myometrium to the effects of oxytocic agents; (*c*) prolonging the life of the corpus luteum and stimulating its secretion; (*d*) acting synergically with progesterone in the development of the endo-

[1] HAMILTON, H. F., *Edinburgh M. J.*, **57**, 1, 1950; "Toxemias of Pregnancy," Ciba Foundation Symposium, J. & A. Churchill, London, 1950, p. 135.

metrium; (*e*) stimulating growth of the mammary gland in preparation for lactation.

Chorionic gonadotrophin found in the urine of pregnant women is produced by the placenta. It increases considerably between the fortieth and eightieth day of pregnancy and then falls to a lower level, where it remains (Fig. 285). The part it plays in the adjustments of pregnancy is still unknown. Pregnandiol is a metabolic product of progesterone secreted by the corpus luteum of pregnancy and in part by the placenta.

The activity of the corpus luteum persists up to the fourth or fifth month of pregnancy, the corpus luteum then gradually undergoing involution. It secretes progesterone, which fulfills the following functions: (*a*) progestational development of the endometrium in the second half of the menstrual cycle, characterized by glandular development and secretion; (*b*) continuance of pregnancy by acting on the placenta; (*c*) inhibition of ovulation and menstrual cycles during pregnancy; (*d*) inhibition of the tonus of the uterus; (*e*) stimulation of alveolar development in the mammary gland after estrogens have stimulated mammary growth; (*f*) softening and relaxation of pelvic ligaments (see "Functions of the corpus luteum," Chap. 59).

Urinary excretion of corticosteroids increases during pregnancy. It shows two peaks, one around the eightieth day, the other around the two hundred twentieth day (Venning, 1946).

The placental hormones. The placenta produces estrogens, chorionic gonadotrophin, adrenocorticotrophin, and small amounts of progesterone.[1] It is the main source of estrogens in pregnancy, as shown by the fact that these substances continue to be produced at an increasing rate in pregnant women who have been castrated during the second or third month of pregnancy.

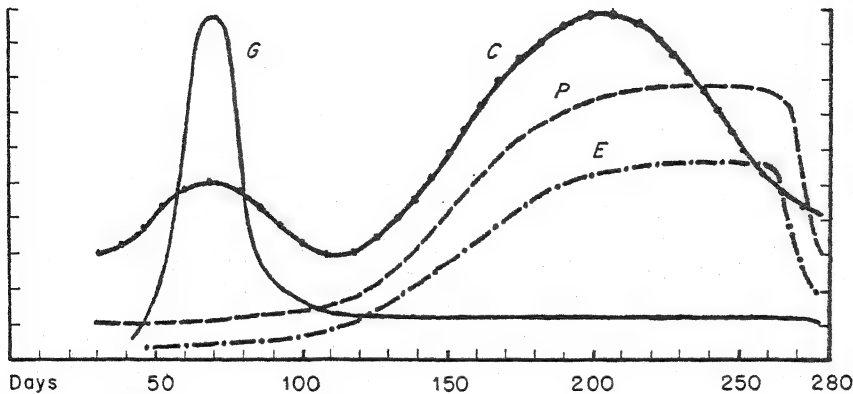



Fig. 285. Urinary excretion of hormones and chorionic gonadotrophin in blood during pregnancy. The peak of each curve corresponds approximately to the following figures: *G*, chorionic gonadotrophin, 500 IU/cc. of serum; *C*, corticosteroids, 200 glycogenic units; *P*, pregnandiol, 40 to 100 mg./day; *E*, estrogens, 15 to 45 mg./day (mostly estriol, 1 to 3 mg. of estrone). (*Data from Venning, Albert, etc.*)

The chorionic gonadotrophin in the urine of pregnant women differs from chorionic gonadotrophin found in the blood serum of pregnant mares. It differs also from anterior hypophyseal gonadotrophins; it is produced by the placenta, where it is found in large quantities, and it continues to be excreted in the urine after expulsion of the fetus if the placenta has been retained. In cases of placental tumors (hydatiform mole and chorioepithelioma) large amounts of gonadotrophins are eliminated in the urine. The gonadotrophin content of the hypophysis diminishes considerably in pregnant women, possibly owing to the great amount of estrogens poured into the circulation by the placenta.

Placental hormones maintain the activity of the corpus luteum. Collip, Selye, and Thomson showed that the corpora lutea persisted in rats after the fetuses were removed between the ninth and thirteenth day of pregnancy, if the placentas were left in the uterus. If the placentas were also removed, the corpora lutea retrogressed rapidly.

Progesterone has been extracted from the

[1] Newton, W. H., *Physiol. Rev.*, **18**, 419, 1938; Newton, W. H., in Allen, "Sex and Internal Secretions," Baillière, London, 1939, p. 720.

placenta, which seems to produce small amounts of this hormone. Ovariectomy causes abortion in rats owing to the suppression of the corpora lutea. If toward the middle of pregnancy the ovaries are extirpated and all the fetuses except one are removed, leaving all the placentas, the single remaining fetus is carried on to full term and is born normally, as if the placentas secreted progesterone (Haterius). In that case, the placentas are expelled from the uterus at the end of parturition, and after delivery milk is secreted normally.

THE PHYSIOLOGY OF THE FETUS[1]

Prenatal growth. The fertilized egg is approximately 0.2 mm. in diameter. It soon divides and the organism grows very rapidly during its embryonic and fetal life. The rate of growth is highest in the early stages and diminishes progressively up to the end of prenatal life.

The newborn child usually weighs from 3 to 3.5 kg. (6.6 to 7.7 lb.), smaller infants weighing as little as 2.4 kg. (5 lb.) and larger ones as much as 4.5 kg. (10 lb.); rarely even smaller or larger infants are born and survive. Males are on an average heavier (3.4 kg. = 7.4 lb.) than females (3.3 kg. = 7.2 lb.). The infant at birth measures 48 to 52 cm. (18 to 20 in.).

The fetal circulation. The circulation of blood in the fetus differs considerably from the circulation after birth. The differences are due mainly to the existence of the placental circuit and the absence of pulmonary respiration. Gaseous interchange in the fetus takes place through the placenta. The circulatory system of the fetus is a closed circuit, completely independent of the maternal circuit, and in no place or stage of development is there any mixture of maternal and fetal blood.

The fetal circulation has been studied *in vivo* in the lamb[2] by means of roentgencinematography after the injection of opaque substances. Oxygenated blood goes from the placenta to the fetus in the umbilical vein. On reaching the liver the greater part passes through the ductus venosus into the inferior vena cava; the rest is distributed in blood vessels of two-thirds of the liver (the other third receives blood from the portal vein). The blood leaving the liver passes through the suprahepatic veins into the inferior vena cava. The arterial blood from the placenta is therefore mixed with venous blood brought from the abdomen and lower limbs by the portal vein and the vena cava.

The blood from the inferior vena cava enters the right auricle, where it impinges on the eustachian valve (crista dividens) and is divided into two currents. The main current is directed through the foramen ovale of the interauricular septum into the left auricle. Here it is mixed with the small amount of blood coming from the pulmonary veins and passes into the left ventricle, from which it is pumped into the aorta. The aorta gives out the coronary arteries to the heart and the arteries going to the head, which thus receive blood with the highest oxygen partial pressure circulating in the fetus. This pressure is, however, lower than the oxygen pressure in arterial blood after birth. The aorta also gives out branches for the different viscera and ends in two large umbilical arteries which carry blood from the fetus to the placenta.

The smaller of the two currents in the right auricle is mixed with venous blood from the head and upper limbs brought by the superior vena cava and blood from the coronary sinus. It passes into the right ventricle and is driven into the pulmonary artery. The greater part of this blood passes through the ductus arteriosus into the aorta; a small part goes to the lungs in the branches of the pulmonary arteries.

Radiographic examination has shown that on tying the umbilical cord considerable changes rapidly take place. First the umbilical veins are closed in the placenta, and the opening into the ductus venosus is occluded by the contraction of a sphincter. The eustachian valve closes the foramen ovale, and the ductus arteriosus is closed by the contraction of its own muscle fibers, which takes place a few minutes after the umbilical cord has been tied. All the blood from the right ventricle now passes into the lungs, and the pulmonary circuit is established.

Patten[1] maintains that the change from the prenatal to the postnatal type of circulation does not occur suddenly; he bases his conclusions

[1] Barcroft, J., "Researches on Prenatal Life," Blackwell, Oxford, 1946; Preyer, J., "Spezielle Physiologie des Embryo," Leipzig, 1885; Windle, F., "Physiology of the Fetus," Saunders, Philadelphia, 1940.

[2] Barclay, A. E., K. J. Franklin, and M. M. L. Pritchard, "The Fetal Circulation and Cardiovascular Systems," Blackwell, Oxford, 1944.

[1] Patten, J., *Am. Heart J.*, **6**, 192, 1930; *Am. J. Anat.*, **98,** 19, 1931.

on anatomical studies. Pereira,[1] however, having made comparative studies of the duration of the different phases of the cardiac cycle in fetuses immediately before and after birth, found significant evidence that the change takes place suddenly or within a very short time.

Fetal respiration. The newly implanted embryo, in its initial stages, breathes in the same way as other cells of the organism, *i.e.*, by absorbing oxygen from, and eliminating CO_2 into, the interstitial fluid of the uterine mucosa. Differences in the respective partial pressures within and without the cells are the cause of this gaseous exchange. At a later stage of development the fetus performs its gaseous interchange (external respiration) through the placenta, which receives fetal venous blood through the umbilical artery and sends arterial blood through the umbilical vein.

Blood gases have been studied in the human fetus immediately after birth and in the fetuses of sheep and goats during the course of pregnancy.

The blood flow through the uterus increases throughout pregnancy, but at a certain moment the fetus grows at a quicker pace than the placenta; therefore, in spite of the increase in placental circulation, during the second half of pregnancy, maternal blood leaving the placenta is increasingly less saturated with oxygen. The oxygen saturation of venous blood from the nonpregnant uterus is about 65 per cent; this figure is reduced to 20 per cent in advanced pregnancy.[2]

There is a significant difference between the oxygen partial pressures of arterial and venous maternal and fetal blood, as is seen in the following figures given by Barcroft:

	O_2 Saturation, Per Cent
Maternal circulation:	
Arterial blood to the uterus	95
Venous blood from the uterus	60
Fetal circulation:	
Venous blood in umbilical artery to the placenta	15 to 45
Arterial blood in umbilical vein from the placenta	70 to 80

[1] PEREIRA, J. C., "Estudios Fonocardiográficos y Hemodinámicos en Embarazadas y en Fetos," Buenos Aires, 1939.

[2] BARCROFT, J., *Physiol. Rev.*, **16**, 103, 1936; NEWTON, W. H., in Evans, "Recent Advances in Physiology," 7th ed., J. & A. Churchill, London, 1949.

The gaseous interchange of O_2 and CO_2 through the placenta can therefore be explained by the differences in partial pressures alone.

Nevertheless in fetuses of goats some paradoxical facts have been observed; *e.g.*, 80 per cent HbO_2 can be found in fetal arterial blood, while maternal blood in the placental sinuses is only 66 per cent saturated. Active secretion of oxygen from the mother to the fetus has been postulated, but this is an unnecessary hypothesis. The oxygen dissociation curve of fetal hemoglobin differs from that of adult hemoglobin; the affinity for O_2 of fetal hemoglobin is greater than that of adult hemoglobin, especially at HbO_2 percentages below 70. The difference is sufficient to explain the paradoxical results observed. Physicochemical conditions of fetal blood are such that CO_2 is set free more readily than in adult blood.

Arterial fetal blood reaches the tissues with a low oxygen content because it is mixed with venous blood, as it is distributed to the fetal portal system and through the ductus venosus and inferior vena cava to the heart (see "The fetal circulation," page 700). The normal condition of the fetus is therefore one of relative anoxia when compared with that of the child after birth.

After delivery, the placental circulation ceases suddenly and pulmonary respiration must be established immediately.

During fetal life there are a few rudimentary respiratory movements, which are inhibited during labor. After birth two factors initiate respiratory movements: (*a*) *asphyxia*, *i.e.*, a decrease in O_2 tension and an increase in CO_2 tension; (*b*) *reflexes* provoked by cold and by tactile or painful stimuli on the skin. The importance of asphyxia can be demonstrated by tying the uterine artery before birth; this procedure provokes respiratory movements in the fetus. Respiratory reflexes can be initiated by opening the uterus, but these do not take place if the fetus is submerged in saline solution at body temperature.

Prolonged and difficult labor increases fetal anoxia. Anesthetics, given to the mother to decrease pain, pass into the fetus and depress the excitability of the respiratory center. In these conditions the fetus frequently does not breathe after birth and enters into what is known as asphyxia neonatorum (fetal asphyxia). Classic remedies for fetal asphyxia are slapping the

baby, or submerging it alternately in warm and cold water to provoke respiratory reflexes by stimulation of the skin, or manual artificial respiration. In cases that are not serious, these methods are sufficient to bring about respiration.

A more efficacious method consists in rhythmic insufflation of the lungs with a mixture of 95 per cent O_2 and 5 per cent CO_2, by means of a rubber catheter introduced into the trachea. The pressure should not rise above 15 cm. of water; otherwise the delicate lung tissues may be damaged. Lobeline is sometimes injected into the umbilical vein in order to increase the excitability of the respiratory center.

Fetal atelectasia does not disappear completely unless the baby makes deep inspiratory movements, which are more efficient than insufflation for expanding the lungs. It is therefore advisable to make premature babies and those with severe asphyxia neonatorum breathe 95 per cent O_2 and 5 per cent CO_2 at intervals to stimulate the respiratory center and expand the lung by repeated deep inspiration, thus diminishing the risk of pulmonary infection.

PARTURITION

The expulsion of the fetus and its membranes is brought about by contractions of the muscle fibers of the uterus. The fetus is adapted to the genital canal through which it must pass by the adoption of adequate positions, and the canal is modified in the act of parturition in order to facilitate the passage of the fetus (Sellheim). Three stages can be distinguished in parturition: (*a*) dilatation of the cervix; (*b*) expulsion of the fetus; (*c*) expulsion of the placenta.[1] After the fetus and placenta have been evacuated the uterus enters into tonic contraction, the uterine cavity disappears, and the vascular sinuses are closed by compression, thus preventing further hemorrhage.

Records of uterine contractions can be obtained by external or internal hysterography. Tocodynamometers are used in external hysterography, placing them on the abdomen so as to record hardening on contraction of the upper, middle, and lower parts of the uterus (Reynolds). The technique of internal hysterography is shown in Fig. 286 (Alvarez and Caldeyro-Barcia).

[1] In rodents the fetuses have been removed leaving the placentas in the uterus. At term the placentas are expelled as in a normal delivery. In these animals, therefore, fetal factors are not essential for the onset of labor and parturition.

During the first two stages of labor (dilatation, and expulsion of the fetus) uterine contractions can be recorded by the pressure they exert on the amniotic fluid.[1] Records thus obtained (Fig. 286*A*) show the following facts: (*a*) the uterus contracts rhythmically; (*b*) the strength of contraction, measured by the increase in amniotic pressure, is 30 to 60 mm. Hg; (*c*) basal tonus, *i.e.*, amniotic pressure between contractions, is approximately 10 mm. Hg (intraperitoneal pressure is taken as the base line); (*d*) frequency of contraction varies between two and five every 10 min.; (*e*) each contraction lasts 60 to 120 sec. As labor advances, the strength, duration, and frequency of contraction increase. During the second stage (expulsion of the fetus) the abdominal muscles and the diaphragm contract and add their effect to the uterine pressure. These contractions are provoked by the entry of the fetus into the vagina, which evokes in the mother the desire to "push," as when evacuating the rectum. The duration of labor is 12 to 16 hr. in primiparas and diminishes to 6 to 10 hr. and even less in multiparas.

After the birth of the fetus, uterine contractions separate the placenta from the uterus and send it into the genital canal (lower segment, os, vagina); abdominal contractions are evoked, or the obstetrician exerts manual pressure on the lower abdomen, and the placenta is expelled. Uterine contractions during this stage can be registered by introducing a needle into the umbilical vein and connecting it to a mercury manometer. Changes in placental blood pressure are due to contraction and relaxation of the uterus. Uterine contractions during this stage follow without interruption, and with the same frequency and strength, the contractions of the previous stages (Fig. 286*B*). The first contractions after birth of the fetus are painless, and they were not known to exist before it was possible to register them. For this reason the 5 to 15 min. following delivery of the fetus was thought to be a period of uterine quiescence (physiologic rest period).

Optimum conditions in normal labor exist when (*a*) uterine contractions are strong (over 24 mm. Hg); (*b*) contraction of the fundus

[1] Alvarez, H., and R. Caldeyro-Barcia, *Surg., Gynec. & Obst.*, **91**, 1 and 641, 1950.

predominates over contractions of the middle and lower segments; (*c*) there is good synchronization between the different parts of the uterus; (*d*) rhythm and strength of contractions are regular; (*e*) the amniotic-fluid pressure falls between contractions to the normal tonus level (10 mm. Hg).

Registration of uterine contractions by means of four microballoons connected to electrical manometers has revealed the existence of two pacemakers, one on each side near the opening of the fallopian tubes. Contractions arise in these pacemakers and spread at a velocity of 1 to 2 cm./sec., the whole uterus being covered in 10 to 20 sec. One of the pacemakers may predominate or predominance may alternate between the two pacemakers. Contractions are stronger and more prolonged in the fundus and upper part of the uterus than in the lower segment and cervix; these latter parts are thus progressively dilated. Labor does not progress satisfactorily if the different parts contract asynchronously, or when contractions arise in the lower parts of the uterus.[1]

[1] Caldeyro-Barcia, R., and H. Alvarez, *J. Obst. Gynaec. Brit. Empire*, **49**, 646, 1952; First Congress on Sterility and Fertility, New York, 1953.

Uterine contractions are not painful; pain in labor is due to: (*a*) distention of the cervix (dilatation); (*b*) distention of the perineum and vulva (expulsion); (*c*) sustained abnormal hypertonic contraction causing uterine ischemia.[1]

A

B

mm.Hg

70

60

50

40

30

20

10

0

min.

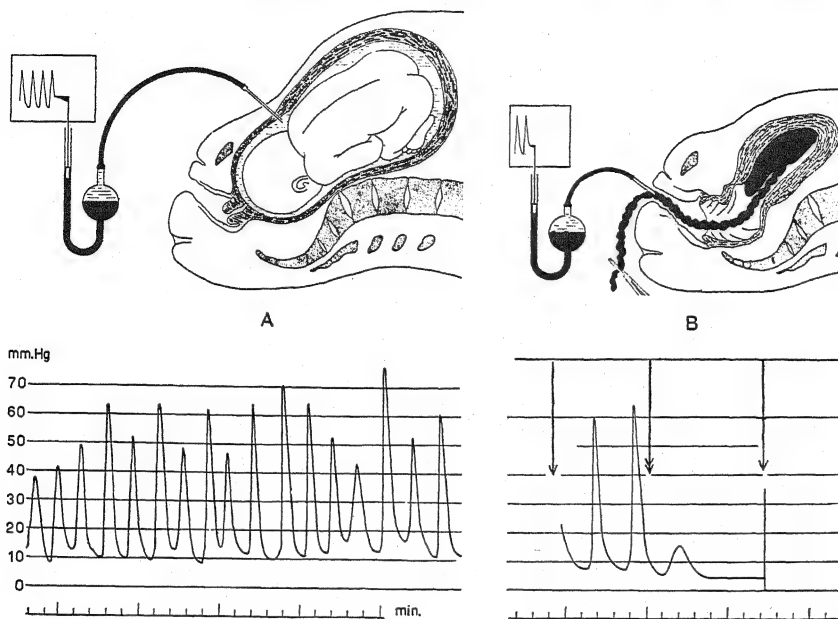



Fig. 286. Registration of uterine contractions. *A*, during labor the amniotic fluid pressure is recorded by a mercury manometer; the record was taken during the period of dilatation. *B*, during the expulsion of the placenta, placental blood pressure is recorded. The first arrow marks the passage of the placenta into the genital canal; the third arrow, manual compression and expulsion of the placenta. (*Alvarez and Caldeyro-Barcia.*)

The cause of the onset of labor after 40 weeks' gestation is still unknown. At term it is possible to initiate parturition by sensitizing the uterus with estrogens and then injecting oxytocic agents, or by dilating the cervix. The appearance of oxytocic substances in the blood at the end of gestation has been reported (Almagia, Fontes), but these substances have also been found in the blood of nonpregnant women (Figueroa-Casas).

The role of the neurohypophysis in parturition has been discussed in Chap. 52.

The uterus is innervated by sympathetic and parasympathetic fibers. Stimulation of these fibers causes the uterus to contract. Reflex contraction of the uterus is provoked by dilatation of the cervix and by stimulation of the nipples (sucking), bladder, or colon. The innervation of the uterus is not essential for parturition.

[1] Moir, J. C., *J. Obst. Gynaec. Brit. Empire*, **46**, 400, 1952; Alvarez, H., and R. Caldeyro-Barcia, *Surg. Gynaec. & Obst.*, **91**, 1, 1950.

Normal labor and delivery have been observed in bitches after denervation of the uterus (Rein, 1880) and after section of the spinal cord at the level of the thorax or destruction of the lumbar and sacral segments of the cord (Goltz, 1874). Parturition has been observed in women suffering from paraplegia (paralysis of the lower half of the body) and in a few women who had suffered transverse section of the spinal cord. Delivery has been observed after the death of the mother, and cases have been reported in which a full uterus contracted after having been surgically removed, the cervix dilated, and the contents were expelled. Destruction of the lumbar and sacral segments of the spinal cord, however, causes difficulties in labor, especially during the expulsive stage. The existence of a spinal center at the level of the tenth thoracic to second lumbar segments, which controls the movements of labor, has been postulated.

BIBLIOGRAPHY

Ahlmark, A., *et al.*, Histaminase in Normal and Pathological Pregnancy, in "Toxemias of Pregnancy," Ciba Foundation Symposium, J. & A. Churchill, London, 1950.

"The Biochemistry of Fertilization of the Gametes," Symposium No. 7 of the Biochemical Society, Cambridge, New York, 1951.

Chang, M. C., and G. Pincus, Physiology of Fertilization in Mammals, *Physiol. Rev.*, **31**, 1, 1951.

Courrier, R., "Endocrinologie de la gestation," Masson et Cie, Paris, 1945.

Farris, E. J., "Human Fertility," Author's Press, Palisades Park, N. J., 1950.

Hotchkiss, R. S., Diagnosis of Testicular Deficiencies, *Bull. New York Acad. Med.*, **27**, 510, 1951.

MacLeod, J., Male Factor in Fertility and Infertility, *Fertil. and Steril.*, **1**, 347, 1950.

Mann, T., and C. Lutvak-Mann, Secretory Functions of Male Accessory Organs of Reproduction in Mammals, *Physiol. Rev.*, **31**, 27, 1951.

Pundel, J. P., and Z. Jaworski, "Gestation et cytologie vaginale," Masson et Cie, Paris, 1952.

Reynolds, S. R. M., Determinants of Uterine Growth and Activity, *Physiol. Rev.*, **31**, 244, 1951.

Reynolds, S. R. M., "Physiological Basis of Gynecology and Obstetrics," Charles C Thomas, Springfield, Ill., 1952.

Reynolds, S. R. M., "Physiology of the Uterus," 2d ed., Hoeber, New York, 1949.

"Toxemias of Pregnancy," Ciba Foundation Symposium, J. & A. Churchill, London, 1950.

Whitelaw, G. P., and R. H. Smithwick, Some Secondary Effects of Sympathectomy, *New England J. Med.*, **245**, 121, 1951.

CHAPTER 62

Lactation

THE FUNCTION OF THE MAMMARY GLAND

The mammary gland is an essential constituent of the reproductive system in all mammals. It secretes milk, which is the natural food of the newborn and also an important foodstuff in adult human nutrition.

Four aspects can be considered when studying the function of the mammary gland: (*a*) the development of the gland; (*b*) galactogenesis, *i.e.*, the processes that originate the secretion of milk; (*c*) galactopoiesis, *i.e.*, the processes that maintain milk secretion; (*d*) the evacuation of milk from the gland.

The Formation and Development of the Mammary Gland

The mammary glands arise from the ectoderm. Lateral to the mid-line on the ventral aspect of the body are the so-called "mammary lines" on which there appear a series of growths that give rise to the mammae. Most mammals have several pairs of mammary glands, but in man only one pair develops, except in rare cases in which small accessory mammae are found.

There are two stages in the development of the mammary glands: at first the ducts grow and ramify; later, during pregnancy, alveoli are formed at the ends of the ducts, and the gland is fully developed.

Growth. The mammary gland develops and undergoes involution at certain ages and in certain physiologic conditions.

1. In the newborn of both sexes there is definite activity in the mammary gland, which is due to maternal hormones that have passed through the placenta into the fetus.
2. There is gradual, but not very considerable, mammary growth during childhood. The prepuberal gland consists of only a few rudimentary ducts.
3. At puberty the glands develop considerably in the female, but only slightly and transitorily in the male.
4. At each estrus there is some development, followed by retrogression during diestrus or anestrus. In women there is some retrogression during the first week of the cycle, and later proliferation up to the premenstrual stage. Some women experience tension and discomfort in the mammary glands just before the menses.
5. During pregnancy the mammary glands reach full development. In the course of the first half of pregnancy not only do the ducts grow, but also the alveoli and lobes are formed; during the second half the gland becomes engorged with the products of secretion.
6. In animals showing the phenomenon of pseudopregnancy at each unfertilized cycle, the mammae develop completely as in pregnancy (bitch, weasel), or incompletely, reaching only the degree of growth observed halfway through pregnancy (rabbit).
7. The mammary glands of women in lactation show still further growth.

Secretion. The newborn of both sexes frequently secrete a few drops of milk. During the last days of pregnancy and a day or two after delivery, small quantities of a fluid called colostrum are secreted. After this interval comes the ejection of milk (the "letdown"), and lactation, as it is usually understood, begins. The mammary glands secrete increasingly large quantities of milk during 6 to 9 months. Later the output diminishes gradually, and it ceases 12 to 18 months after delivery. If the child

continues to be nursed and sucks vigorously, lactation is kept up for a longer time; exceptional cases of lactation lasting 2 and 3 years have been reported.

Involution. If the offspring are not nursed, the mammary gland rapidly undergoes involution, and the same process takes place at whatever stage lactation is suspended. The gland retrogresses and in a few days returns to a condition similar to that which existed before pregnancy. Castration is quickly followed by marked involution. After the menopause the mammary glands retrogress gradually.

The male mammary gland. Slight transitory growth is observed in the male mammary glands at birth and at puberty. Deposition of fat in the mammary region may simulate glandular development, *e.g.*, in some cases of dystrophia adiposogenitalis. Growth of the mammary gland itself (gynecomastia) has been reported in males with certain testicular and corticoadrenal tumors, and in some chronic hepatic diseases, malnutrition, etc.

Hormonal Factors in the Development of the Mammary Gland

Growth and secretion of the mammary gland are stimulated and controlled by several endocrine factors, *i.e.*, internal secretions of the ovary, the corpus luteum, the placenta, and the pars distalis of the hypophysis. The ovary and the placenta secrete estrogens and progesterone, and the hypophysis secretes gonadotrophins, *i.e.*, FSH, LH, and luteotrophin (prolactin). Prolactin has a triple effect: it is luteotrophic, mammotrophic, and lactogenic. The mammae are influenced by adrenal and thyroid secretions and by nutritive conditions (Fig. 287). Neural factors also play a part but are not indispensable for normal development and function.

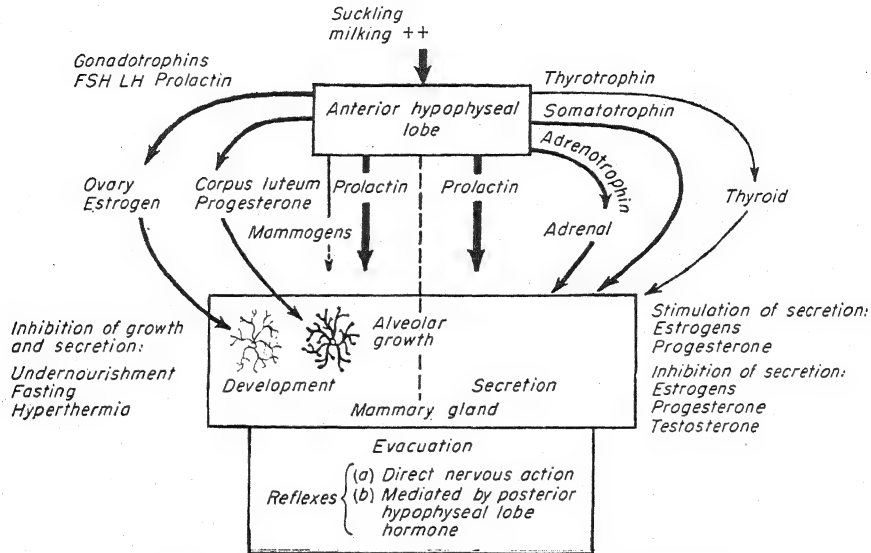



Fig. 287. Diagram of factors that control the secretion of milk.

Estrogens, secreted by the ovary or the placenta, initiate the development of the gland, and progesterone completes it. Hypophyseal function is a necessary factor in both stages.

The role of the ovary. Normal development of the mammary gland is due to the action of the ovary which (*a*) stimulates puberal growth of the gland, (*b*) maintains its development during adult life until the menopause; (*c*) provokes its growth during the first part of pregnancy (later the placenta also takes part in this process).

The importance of the ovary is demonstrated by the effects of castration and restitution of ovarian function. Prepuberal castration prevents the puberal growth of the mammae, and postpuberal castration causes rapid involution. Retrogression of the mammary gland after the menopause is also due to the decrease in ovarian function which occurs in old age. The effects of castration are prevented or controlled by ovarian grafts or the administration of natural

estrogens, such as estrone and estradiol, or of artificial estrogens, such as diethylstilbestrol.

The action of the ovary on the mammary glands is due to the secretion of estrogens which is afterward completed by the hormone of the corpus luteum (progesterone). Injections of these hormones in several species have shown that there are three types of response: (*a*) in many animals (*e.g.*, mice and dogs) estrogens provoke growth of the mammary ducts only, and progesterone also must be injected in order to obtain alveolar development; (*b*) in a few species (*e.g.*, guinea pigs[1] and monkeys) estrogen alone provokes growth of the ducts and alveoli; (*c*) in other species estrogens stimulate the development of the ducts and to a certain extent that of the alveoli; thus in cows and goats there is some alveolar growth and milk secretion, but full normal growth is obtained only if progesterone is also injected (Turner).

Estrogens have a direct effect on the mammary gland and a more important indirect effect which is exerted through the anterior hypophysis and which will be considered in a following paragraph. Direct stimulation of mammary growth by estrogens has been observed in castrated hypophysectomized animals, but it is less marked than in normal animals. Local application of estrogen on the breasts by means of ointments also causes mammary development not only in normal animals but also in hypophysectomized ones. If estrogen is applied to only one mammary gland, it develops and the others remain unchanged. If the dose of estrogen applied locally is high, some growth is also observed in the other glands. Local application of estrogen causes development of the nipple, areola, and mammary gland in women and to a certain extent in men.

Large doses of estrogen cause abnormal development of the ducts with formation of cysts (Gardner).

In certain strains of mice there is a high incidence of cancer of the mammary gland in the female; ovariectomy diminishes this incidence considerably (Loeb).

[1] In male castrated guinea pigs ovarian grafts cause full development of the mammae, similar to that which occurs in pregnancy, and the secretion of milk. In female castrates the effect is less marked. The difference has been attributed to the fact that the male hypophysis secretes continuously large quantities of gonadotrophin, which stimulate maximum follicular growth and estrogen secretion. The female hypophysis secretes more luteinizing hormone than the male; therefore corpora lutea are formed and the estrogenic effect is periodically inhibited.

Cancer of the mammae seldom occurs in the male; treatment with estrogen causes the glands to develop, however, and then many of them become cancerous.[1] Even in noncancerous strains of mice, early treatment with large doses of estrogens provokes the formation of adenocarcinomas of the mammae in almost half the individuals.[2] Moreover some of the cancerigenic substances provoke growth of the ducts and alveoli of the mammae.

Testosterone stimulates growth of the ducts, especially when given with an adequate dose of estrogen. In large doses it has an inhibitory effect and causes involution of the gland. Testosterone treatment is given in cases of cancer of the breast because often it reduces the size of the tumor, diminishes pain, and facilitates the surgical removal of the gland.

The hormone of the corpus luteum is needed for the full development of the mammae, except in the few cases just mentioned. This hormone given in very large doses can provoke alveolar growth, but its maximum effect is observed only after the glands have been sensitized by a previous treatment with estrogen. Associating both hormones in appropriate doses, each one potentiates the action of the other; if the relative doses are not adequate, they oppose each other's action.[3] The importance of the corpus luteum for mammary development can be easily demonstrated in species that show pseudopregnancy in the nonfertilized cycles. The growth of the mammae is parallel to that of the corpus luteum, and premature destruction of the latter causes retrogression in the former (Ancel and Bouin).

The role of the placenta. As pregnancy advances the placenta becomes the most important source of estrogens and gonadotrophin (chorionic), a less important source of progesterone, and a source of small amounts of prolactin and adrenocorticotrophin. Cases have been reported of women who have suffered ovariecto-

[1] Lacassagne, A., *Compt. rend. Acad. d. sc.*, **15**, 630, 1932; *Am. J. Cancer*, **37**, 414, 1939.

[2] Geschickter, C. F., and E. W. Byrnes, *Arch. Path.*, **33**, 334, 1942. Hypophysectomized animals did not respond to estrogens by the formation of tumors in the mammae. It has not been possible to provoke adenocarcinoma in the mammary glands of female monkeys and rabbits by estrogen treatment.

[3] Other substances have the same effect as the hormone of the corpus luteum; taking as 1 the activity of progesterone, that of methyltestosterone equals 1/6, desoxycorticosterone and dehydroandrosterone 1/8, and diethylstilbestrol 1/4 (Mixner and Turner, 1942).

my in the third or fourth month of pregnancy in whom the mammary glands continued to grow and secreted milk after the delivery. In mice the ovaries and fetuses can be removed without interfering with mammary growth if the placentas are left in the uterus, but the mammae undergo rapid involution if the placentas are also extirpated. The mammary development observed in the newborn is probably due to maternal ovarian, hypophyseal, and placental hormones which pass into the fetus during the last stages of pregnancy.

The role of the hypophysis. The pars distalis of the hypophysis plays an important and multiple role in the development of the mammary gland. Gonadotrophin secretion stimulates growth and secretion of the ovary and corpus luteum, *i.e.*, secretion of estrogens and progesterone. Hypophysectomy therefore causes ovarian hypoplasia (in young animals) or atrophy (in adults) and thus hypoplasia or atrophy of the mammary gland.

Three hormones are necessary for the complete development of the mammary gland: (*a*) estrogen; (*b*) progesterone; (*c*) prolactin. Estrogen and progesterone do not provoke mammary growth in hypophysectomized animals, but exert their full effects if a small amount of prolactin is also injected. No other hypophyseal hormone has this effect (Lyons, Nelson). Somatotrophin and adrenocorticotrophin cannot replace prolactin, but they increase the activity of the latter.

Estrogens have a direct effect on the mammary gland, but it is slight or nonexistent in hypophysectomized animals. This peripheral direct effect can be demonstrated by the application of salves or solutions of estrogens on one gland; the nipple, areola, and gland develop. The gland on the other side is not affected, unless large doses of estrogen have been absorbed, in which case it may develop, although in a lesser degree. A smaller effect can be produced in men by the local application of estrogen on the mammary area.

Prolactin has a local activity; if it is injected into a mammary gland, growth and milk secretion are provoked only in that gland, not in others.[1]

Turner and his associates demonstrated that injection of hypophyseal extracts provokes mammary development in hypophysectomized castrated animals. The active principles (two were postulated) were called mammogens. Later only one was considered as different from prolactin, but this mammogen has not been adequately separated from hypophyseal extracts, and its existence is still doubtful.

[1] Lyons, W. R., *Proc. Soc. Exper. Biol. & Med.*, **51**, 308, 1942.

The Secretion of Milk

Two aspects of milk secretion can be considered, *i.e.*, initiation (lactogenesis) and maintenance (galactopoiesis).

The fact that milk is not excreted until after delivery led to the belief that the "letdown" was due to the suppression of an inhibitory factor. Later positive hormonal and nervous factors that initiate and maintain the secretion of milk were demonstrated. The hormonal factors are hormones of (*a*) the pars distalis; (*b*) the sexual glands; (*c*) the adrenals; (*d*) the thyroid.

The action of the hypophysis. The part played by the pars distalis has been demonstrated by the effects of hypophysectomy and of injection of hypophyseal extracts or hormones.

Hypophysectomy in pregnant females provokes abortion in nearly all species,[1] but the secretion of milk that is usually seen after abortion does not occur in these animals. If the hypophysectomy is performed in a lactating female that is nursing her young, the secretion of milk ceases almost immediately; if it is performed shortly after delivery, a slight transitory secretion is sometimes observed.[2]

Daily injections of anterior hypophyseal lobe extract provoke copious milk secretion in a few days if the mammary glands are well developed.[3] In male dogs these extracts produce some secretion of milk if mammary development has first been obtained by estrogen treatment. In the bitch hypophyseal extracts stimulate the secretion of milk even after the ovaries, thyroid, and adrenal medulla have been extirpated.

Riddle obtained a lactogenic hormone from the hypophysis which he called "prolactin."[4] It is a protein with a molecular weight of 32,000 to 35,000, which has been purified and obtained in crystalline form.[5] This substance has the fol-

[1] In the rat hypophysectomy is followed not by abortion but by prolonged pregnancy.

[2] Allen and Wilkes, 1932; Collip, Selye, and Thomson, 1932; Houssay, 1935; Nelson, 1935.

[3] Stricker and Grüter, 1928; Corner, 1930; Houssay, 1935.

[4] It has also been called "galactin" (Turner), "mammotropin" (Lyons), "lactogen" (Turner), and "luteotrophin" (Hisaw and Astwood).

[5] White, A., *Ann. New York Acad. Sc.*, **43**, 341, 1943.

lowing specific effects: (*a*) it provokes milk secretion in fully developed mammary glands; (*b*) in pigeons, it stimulates hypertrophy of the crop glands and the milklike secretion of the crop; (*c*) it maintains the development of the corpus luteum; (*d*) injection of this hormone causes maternal behavior in virgin rats, and broodiness in hens; (*e*) metabolic effects are revealed by changes in the weight of several organs (Riddle *et al.*); (*f*) it provokes secretion of the oviduct in the toad (Allende, 1939).

The lactogenic hormone stimulates epithelial proliferation and desquamation in the crop of pigeons; this produces a caseous mass (the so-called "crop milk") which serves to nourish the young. This response is used for the biological standardization of prolactin. Two methods are used: (*a*) a micromethod, consisting in the injection of the substance to be assayed under the skin near the crop, by which very small amounts of prolactin can be detected such as those which have been found in the blood and urine of women in labor and in the hypophysis of small animals; (*b*) a macromethod, consisting in determining the increase in the weight of the crop after injecting the substance to be assayed into the pectoral muscles of the bird. The international unit of prolactin is the specific activity of 0.1 mg. of the international standard preparation.

Prolactin alone does not provoke milk secretion in hypophysectomized animals (*e.g.*, in guinea pigs). Milk secretion can be induced by the injection of total hypophyseal extract, or by injecting adrenocorticotrophin, or adrenocorticosteroids simultaneously with prolactin.[1] Milk secretion is due to the coordinated activity of a hormonal complex, in which prolactin is the main and indispensable lactogenic factor (Folley and Young).

Hypophyseal extracts increase the milk production of cows by 20 per cent when it begins to diminish, but not when it is at a maximum.[2] Somatotrophin has a similar effect. Total extract is more efficacious than prolactin, and its effect is apparently due to the combined activity of several hormones. The high cost of the treatment prevents its use to increase the commercial yield of dairy cows. In women no satisfactory results have so far been obtained.

[1] Nelson, W. O., R. Gaunt, and M. Schweizer, *Endocrinology*, 33, 325, 1943.

[2] Young, F. C., and S. J. Folley, *J. Endocrinol.*, 4, 194, 1944.

In human hypophyseal insufficiency there is sexual infantilism; therefore, owing to ovarian hypoplasia, the mammae do not grow. In acromegaly, on the other hand, lactation is prolonged, and in one reported case it lasted 5 years. Prolactin seems to be secreted by the acidophil cells of the anterohypophysis. The hypophyseal content of prolactin can be increased by injecting estrogens immediately after delivery, and by suckling (Turner).

Removal of the posterior lobe of the hypophysis in rats (Smith, 1932) and dogs (Houssay, 1935) does not cause any change in the secretion of milk.

The effect of estrogens. Natural and synthetic estrogens stimulate mammary growth and increase the lactogenic hormone content of the hypophysis, but they have inhibitory effects on milk secretion in large doses; their activity is increased by an adequate dose of progesterone.

In certain animals estrogens alone can develop the mammae and stimulate milk secretion if the hypophysis is intact. Thus diethylstilbestrol and hexestrol produce alveolar growth and copious milk secretion in virgin and adult goats, sheep, and cows.[1] Mammary development thus obtained differs cytologically from the normal development of pregnancy (Turner).

The main inhibitory effect of estrogens is apparently exerted by directly antagonizing the effect of prolactin (Nelson), but it may also diminish prolactin secretion by the hypophysis. Folley and Young point out that there are two thresholds: if estrogens increase up to a certain level, the secretory threshold is reached; a further increase of estrogens will reach the inhibitory threshold. Ovariectomy does not suppress milk secretion in lactating cows; frequently lactation is prolonged beyond the usual time for it to cease.

The action of the adrenals. Adrenalectomy diminishes but does not suppress the secretion of milk in most animals. The secretion is slightly improved by treating adrenalectomized animals with sodium chloride, and is restored to the normal level by corticoadrenal hormone. Corticoadrenal steroids with oxygen on C^{11}, *i.e.*, those which are active in protein and carbohydrate metabolism, are particularly efficacious. Desoxycorticosterone, which acts on salt and water metabolism, has little effect on the milk secretion of rats (Gaunt) and inhibits it in the guinea

[1] Folley, J. H., J. Hammond, and A. S. Parker, *J. Endocrinol.*, 4, 1, 1944.

pig. On the other hand, Verzár[1] has used it with success in the treatment of pregnant adrenalectomized cats, in which the fetuses were carried to term and, in some cases, nursed.

The lactogenic hormone does not produce milk secretion in hypophysectomized guinea pigs unless it is associated with corticoadrenotrophin or corticoadrenal hormones or extracts.[2] It is not yet possible to state whether disturbances in milk secretion following adrenalectomy are due to general metabolic disturbances or to the lack of a specific lactogenic hormone. There is no proof of the existence of a "corticolactin," which has been postulated by some workers.

The action of the thyroid. Thyroidectomy does not prevent the development of the mammae that occurs during pregnancy nor the secretion of milk following delivery, but the amount secreted is definitely subnormal. Thyroidectomy reduces the amount of milk secreted by lactating animals. The administration of thyroid preparations, thyroxine, or iodized proteins increases milk secretion. Diethylstilbestrol does not cause mammary growth in thyroidectomized cows unless thyroid insufficiency is improved by treatment with iodized proteins.[3]

The role of the nervous system. Stimulation of nerves going to the mammary gland produces only vasomotor effects and the expulsion of a little milk contained in the alveoli and ducts of the gland; it has no secretory effects. The dispensability of mammary innervation for the normal growth of the gland during pregnancy and milk secretion after delivery has been demonstrated in several ways:

1. Extirpation of the sympathetic chains or denervation of the mammary gland does not prevent milk secretion.
2. Milk secretion is not prevented by the destruction of the spinal cord from the third thoracic segment down (Goltz and Ewald, 1886). A case of a woman with section of the spinal cord at the level of the sixth thoracic segment, who delivered and nursed a child, has been reported.
3. A grafted mammary gland develops normally during pregnancy and secretes milk.
4. The pygopagus twins Theresa and Josepha Blazek had separate genital organs and nervous systems, but their circulations were anastomosed. One of them became pregnant and delivered a child, but the mammary glands of both of them developed and, after delivery, secreted milk.

On the other hand manipulation of the mammae, especially suction, can eventually provoke mammary development and milk secretion sufficient to nurse sucklings. This has been observed in nonpregnant and virgin females and exceptionally in males in many species (man, dog, goat, rats, mice, etc.). The following facts are evidence that suckling initiates a reflex which increases secretion of galactopoietic hormones by the hypophysis: (*a*) suckling does not maintain milk secretion of lactating animals after hypophysectomy or adrenalectomy; (*b*) suckling stimulates the secretion of prolactin; (*c*) it causes loss of granules in the acidophil cells, and it prevents the appearance of castration cells in the hypophysis after ovariectomy (Desclin).

Suckling plays an important part in other phenomena: (*a*) it provokes uterine contractions after parturition, and accelerates normal uterine involution during the puerperium; (*b*) it prolongs the life of the corpus luteum (corpus luteum of lactation), thus inhibiting ovulation and the menstrual cycle.

Suckling or milking is necessary for the maintenance of normal milk secretion. Soon after they are discontinued, the secretion ceases and involution of the mammary gland commences. Frequent suckling or milking, with complete evacuation of the gland, is the most powerful known factor for increasing milk secretion when it begins, or declines, or is scarce. Habitual incomplete evacuation causes a decrease in milk secretion and, eventually, involution of the gland.

Suction may act by initiating reflexes from the nipple, or simply by preventing dilatation of the gland due to the milk retained. According to Selye, suction provokes reflex discharges which act on the hypophysis. This is demonstrated by tying the ducts so as prevent the evacuation of the gland. As long as suction is maintained, secretion continues; but soon after suction is discontinued, or if the nipples are cut, the gland rapidly undergoes involution. Undoubtedly there is a local factor acting on the

[1] Verzár, F., *Helvet. Physiol. et Pharmacol. Acta*, **1**, 385, 1943.

[2] Nelson, W. O., R. Gaunt, and M. Schweizer, *Endocrinology*, **33**, 325, 1943.

[3] Petersen, W. E., *et al.*, *Proc. Soc. Exper. Biol. & Med.*, **57**, 332, 1944.

gland, because if one gland is not sucked it suffers involution to a certain extent, while the others which are sucked remain fully developed and continue to secrete. Involution of the unsucked gland is nevertheless more gradual than the involution that occurs when no gland is sucked. This is probably due to the secretion of prolactin stimulated by sucking, because injections of prolactin delay involution of unsucked glands.

The flow of milk that occurs after parturition has been attributed to (*a*) the delivery of the fetus; (*b*) expulsion of the placenta; (*c*) ceasing of distention of the uterus; (*d*) neurohypophyseal hormones. It does not occur if the hypophysis is extirpated. In rats, if the placentas remain in the uterus after the fetuses have been removed, milk flow does not begin until the placentas have been expelled. The mechanism of this phenomenon is not yet well understood. It has been suggested it is due to (*a*) ceasing of inhibition of prolactin by progesterone (Turner); (*b*) ceasing of the inhibitory effect of estrogens on the action of prolactin on the mammary gland (Nelson); (*c*) the decrease in estrogen from the inhibitory threshold level to the stimulatory threshold level (Folley).

Other factors. Nutritive conditions are important for the development of the mammary glands, which grow less in response to estrogens in undernourished or fasting animals. Progesterone has little effect on animals submitted to high temperatures. It is common knowledge that emotional shock causes a decrease in milk secretion. Acute and chronic infections have unfavorable effects on the secretion of milk.

The Emptying of the Mammary Gland

The evacuation of the mammary gland is of importance not only for the actual supply of milk to the offspring but also for the continuance of secretion and the prevention of involution. Complete evacuation increases the yield of milk and prolongs lactation; it has therefore been the subject of much study from the point of view of dairy farming.

The emptying of the mammae is not only the passive result of sucking or milking; it is accompanied by contraction of the glandular ducts and controlled by hormonal and nervous factors.

Psychic phenomena exert a definite influence on the evacuation of the mammae. A cow will "hide her milk" if she is disturbed by unfamiliar noises or is frightened, or if the milker is changed. The presence of the calf, on the contrary, produces a more copious flow, and suction by the calf at the end of milking will stimulate further secretion. Conditioned reflexes that facilitate milking can also be provoked by preliminary manipulation of the mammae such as is performed in washing the udder. In nursing women the presence of the child, or hearing it cry, may provoke the flow of milk. When the child sucks at one breast, frequently a little milk flows from the other.

The evacuation of milk is a reflex. The afferent paths are not yet known, but impulses reach hypothalamic centers and are relayed to the neurohypophysis, which secretes hormones provoking contraction of the alveoli and ducts of the mammary glands.

Injection of posterior hypophyseal lobe extract or oxytocin rapidly provokes an abundant flow of milk if the main duct is catheterized or the nipple has been cut off.

Moreover, if a cow or goat that has been milked dry is injected with posterior lobe extract or oxytocin, on further milking it will yield a new amount of milk rich in fat.[1] Oxytocin causes a rise in pressure in the milk ducts, and histological examination shows the almost complete evacuation of the gland. Oxytocin has about five times the activity of purified vasopressin.[2]

Suckling and milking provoke contraction of the smooth muscles of the ducts and the myoepithelial cells of the acini, and the intramammary pressure rises. Stimulation of the nerves to the mammary gland does not produce this effect, so a direct nervous action is still to be demonstrated. A humoral factor is released during milking, because blood collected from a gland which is being milked, if perfused through another gland, will cause this one to evacuate its milk.[3]

The hypothalamic-neurohypophyseal link in the mechanism of milk evacuation is demonstrated by the following facts: Electrical stimulation of the supraoptic nucleus (in ewes)[4] or of

[1] Ott and Scott, 1910–1921; Schüfer *et al.*, 1911–1913; Hammond, 1912; Houssay *et al.*, 1913–1914.

[2] Petersen, W. E., *et al.*, *J. Dairy Sc.*, **24,** 211 and 225, 1941; **27,** 449, 1944.

[3] Petersen, W. E., and T. M. Lukwick, *Federation Proc.*, **1,** 66, 1942: Peeters, G., *et al.*, *Arch. internat. de pharmacodyn. et de thérab.*, **75,** 85, 1947; **79,** 113, 1949.

[4] Anderson, B., *Acta physiol. Scandinav.*, **23, 1,** 8 and 24, 1951; **25,** 212, 1952.

the supraopticohypophyseal tract (in rabbits)[1] provokes evacuation of milk, even after denervation of the mammary gland. Hypertonic salt solution injected into the carotid has the same effect. The antidiuretic hormone is released at the same time as the milk-evacuating factor, so that urinary secretion diminishes and an antidiuretic substance appears in the urine. Section of the pituitary stalk in nursing females is followed by the death of the offspring, probably owing to deficient evacuation of milk; if oxytocin is injected into the female, the offspring survive.

Table 79. The Chemical Composition of Mammary Secretion

Mammary secretion	*Constituent, gm. per 100 cc.*							
	Water	*Protein*	*Casein*	*Lact-albumin*	*Lactose*	*Fat*	*Ash*	*Calories*
Human milk..........	88.5	1.2	0.5	0.8	6.8	4.0	0.2	67
Cow's milk...........	87.1	3.5	2.9	0.6	4.8	3.5	0.7	69
Colostrum...........	....	2.06	...	...	4.6	2.6	0.35	50

MILK

Amount secreted. In women immediately after delivery a small amount of colostrum is secreted; the flow of milk does not begin till 2 or 3 days later. The amount of milk secreted increases gradually up to the sixth or ninth month, and then diminishes until it ceases 12 to 18 months after delivery, but it may last longer if vigorous suction is continued. The total daily amount reaches a maximum of 1,000 to 1,500 cc., but secretions of 2.5 and 3 liters daily, without any harmful consequences to the mother, have been reported. Women who nurse two babies secrete more milk than those nursing only one.

Composition and nutritive value. Milk is a dispersion of corpuscles, micellae, molecules, and ions. The color and opacity of milk are due to fine fat droplets which it holds in suspension.

The specific gravity of milk is seldom below 1.028 or above 1.034. The freezing point is usually between −0.5 and −0.6°C. The pH is 6.6 to 6.8 in human milk and 6.3 to 6.6 in cow's milk. After extraction, if the milk stands at room temperature, the pH is lowered owing to the development of microorganisms, mainly the lactic bacilli which split lactose into lactic acid and sour the milk. The addition of acids coagulates milk when the pH falls to 5.34 in raw milk and to 5.37 in boiled milk. If acid milk is warmed it clots.

The components of milk (Tables 79 and 80) are water, salts, and organic substances.

The principal proteins in milk are (*a*) *caseinogen*, a phosphoprotein which is converted into paracasein by the rennin in gastric juice (paracasein combines with calcium and forms solid casein); (*b*) *lactalbumin;* (*c*) *lactoglobulin*, identical to serum globulin, which includes maternal antibodies; (*d*) a small amount of alcohol-soluble protein. There is also some nonprotein nitrogen (35 mg. per cent).

Table 80. Salts in Milk

Milk	*Salt, mg. per 100 cc.*						
	K	*Ca*	*P*	*Na*	*Mg*	*Cl*	*Fe*
Human....	41	30	13	11	6	36	0.10
Cow's......	150	120	95	50	12	110	0.05

Lactose is the principal sugar in milk, but there are small quantities of other carbohydrates.

The fat in milk is made up mainly (90 per cent) by the glycerides of oleic, myristic, and palmitic acid, and smaller amounts of stearic acid and acids with carbon chains of 4 to 24 carbon atoms. Phosphatides, such as lecithin and kephalin, and sterols are also found.

The main cations in milk are potassium, calcium, and sodium, and the main anions are phosphate and chloride, with small quantities of citrate and lactate.

Vitamins A, C, and most of those in the B groups are found in sufficient quantities to satisfy the needs of the child, but there is not enough vitamin D (Tables 69 and 70, Chap. 49; Table 81).

[1] Cross, B. A., and G. W. Harris, *J. Endocrinol.*, **8**, 148, 1952.

Milk contains antibodies which circulate in the mother's blood and a certain number of enzymes (catalase, oxydase, reductase, etc.). These enzymes are destroyed by heat, and their absence serves as a test of whether the milk has been boiled.

Table 81. Vitamins in Milk

Vitamin	Amount per 100 cc.	
	Cow's	*Human*
Vitamin A, IU	180	290
Vitamin D, IU	2.5	5
Vitamin C, mg	2.0	6.4
Thiamine, mg	0.035	0.013
Riboflavin, mg	0.20	0.04
Niacin, mg	0.1	0.1
Pantothenic acid, mg	0.35	
Biotin, μg	3	
Pyridoxin, mg	0.04	
Choline, mg	15	
Inositol, mg	18	
Folic acid, mg	5	

The composition of milk differs from one species to another. The milk of the offspring's own species is the one best adapted to satisfy its needs. The nutritive value, especially the protein concentration, increases with the speed of growth of the offspring (Table 82). The fact that the composition of the ash of milk is similar to that of the newborn offspring, and not to that of the maternal plasma, is further evidence of the adjustment of the composition of milk to the needs of the offspring.

Milk is the indispensable food of the newborn and the principal foodstuff of the growing child. It is also an excellent food for adults. It is an almost complete food; the only deficiencies are the low iron and vitamin D content. A child that is fed exclusively on milk for too long a time becomes anemic when the iron stored in its body (mainly in the liver and spleen) is exhausted.[1] It is also advisable to add vitamin D to the diet of a milk-fed child, either in the form of cod-liver oil or other fish-liver oils, or by adding it to the milk (see "Vitamin D," Chap. 49).

A 70-kg. adult could obtain the total caloric requirement and an adequate amount of high-class protein from 5 to 6 liters of milk daily. The fat content of such a diet would be too high.

[1] Rats fed exclusively on milk cease to grow and eventually die. Iron and copper must be added to cure the symptoms of deficiency.

Milk not only is an excellent food from a biological point of view, but it has also many advantages from the economic standpoint. The cow puts out in milk 35 per cent of the energy of the food digested and is therefore the most efficient machine for converting energy into animal food of high quality, well adjusted to human needs.

Table 82. Composition of Milk in Relation to Growth of Offspring

Animal	Days taken for offspring to double birth weight	Constituent, gm. %			
		Protein	*Salts*	*Ca*	*Phosphoric acid*
Woman	180	1.6	0.2	0.033	0.047
Mare	60	2.0	0.4	0.124	0.131
Cow	47	3.5	0.7	0.160	0.197
Goat	22	4.3	0.8	0.197	0.284
Sow	14	5.9	0.8	0.249	0.308
Cat	9.5	9.5	1.0		
Bitch	9	7.4	1.3	0.455	0.508
Rabbit	6	10.4	2.5	0.891	0.997

Source: after Porcher, modified.

Colostrum. This is an opalescent, turbid fluid which has in suspension fatty droplets and multinucleated cells filled with fat. It has 2 to 8 per cent protein and coagulates when heated. It has less calories, fat, and carbohydrate than milk, and more protein; it is also more alkaline and has more vitamin A, thiamine, riboflavin, and ascorbic acid, and less niacin. It does not contain vitamin K.[1] It is popularly considered to have laxative properties (Table 79).

Comparison between human and cow's milk. Cow's milk is commonly used as a food for children and adults. Its composition differs from that of human milk (Tables 79 and 80), and it is not possible to make it identical with the latter simply by dilution and the addition of different substances. The main differences between human and cow's milk are the following: (*a*) human milk has much less casein than cow's milk (about one-eighth as much), but the total protein content has a higher nutritive value per gram because the ratio of casein to lactalbumin is between 1:1 and 2:1, while in cow's milk it is between 4:1 and 8:1 (there is

[1] Escudero, P., *Bol. Acad. nac. de med. de Buenos Aires*, p. 355, 1942.

more lactalbumin per gram of total protein in human than in cow's milk); (*b*) the concentration and physicochemical properties of casein in human milk are such that on coagulating it forms a finer, softer, and more uniform curd, which is more easily digested than the firmer curd formed by cow's milk; (*c*) human milk contains less salts, especially Ca and P, than cow's milk;[1] (*d*) the buffer effect is less in human than in cow's milk; (*e*) human milk contains more lactose than cow's milk; (*f*) human milk contains only one-tenth the amount of volatile fatty acids in cow's milk. The characteristic smell of babies is due to the action of saliva and gastric juice on milk (Arthus).

Origin of milk constituents. The use of isotopes, biochemical studies on mammary-tissue slices *in vitro*, and perfusion of the mammary gland have given valuable information on the processes of milk production.

There is no caseinogen in the blood; it is synthesized by the mammary gland from the amino acids that it takes up from the blood. Part of the milk fat probably comes from the blood; thus a certain amount of the fatty acids in food pass into the milk and may contribute to its taste. If the female is fasting, storage fat is mobilized for the formation of milk fat. The mammary gland, however, also synthesizes fat from acetate, a process which is increased by the action of insulin. Thus the composition of milk fat is, to a certain extent, maintained constant. Lactose is formed from glucose in the blood. A special enzyme and coenzyme convert glucose-1-phosphate into galactose-1-phosphate.[2] Lactose formed in the mammary gland of lactating women may pass into the blood and be excreted in the urine.

The influence of the diet on the composition of milk. An increase in the protein content of the diet increases the amount of milk secreted and its fat content up to a certain limit, but does not modify the protein content. An increase of fat in the diet above 1 gm. per kg. per day does not increase the fat content of milk, but if the diet does not contain enough fat, milk fat diminishes. The composition of the diet has little or no influence on the lactose content of milk.

If the diet is adequate in amount and composition, an increase in the ingestion of food does not increase the output of milk. If the diet is inadequate the female takes from the stores in her body those substances which are needed for the formation of milk, but the milk output diminishes. Thus, if there is deficiency of Ca and P in the diet, the female will remove these elements from her own bones and teeth and keep their concentration in the milk nearly normal for a fairly long time, but eventually they will diminish. Iodine and vitamins in milk are dependent on the maternal intake of these substances. In Chap. 50 the importance of supplementing the diet of pregnant and lactating women with iron, iodine, vitamins, etc., was emphasized.

Excretion of foreign substances in milk. Many substances ingested in the food are excreted in the milk and may give it a peculiar odor or taste. Certain drugs also pass into the milk, *e.g.*, alcohol, belladonna, opium, salicylates, iodides, bromides, saline purges, etc.

[1] There is more K, Na, and Cl per gram of Ca in human milk than in cow's milk.

[2] Caputto, R., *et al.*, *J. Biol. Chem.*, **184**, 333, 1950; Leloir, L. F., *et al.*, *An. Assoc. quim. argent.*, **40**, 228, 1952.

BIBLIOGRAPHY

Centre National pour les Recherches Scientifiques. Colloque International. Mécanisme physiologique de la sécrétion lactée, Paris, 1951.

Cowie, A. T., S. J. Folley, B. A. Cross, G. W. Harris, D. Jacobsohn, and K. C. Richardson, Terminology for Use in Lactational Physiology, *Nature, London*, **168**, 421, 1951.

Folley, S. J., and F. H. Malpress, The Hormonal Control of the Mammary Growth and Lactation, in Pincus and Thimann, "The Hormones," Academic Press, New York, 1948.

Folley, S. J., Aspects of Pituitary–Mammary Gland Relationship, *Recent Progress in Hormone Res.*, **7**, 107, 1952.

Folley, S. J., L. Desclin, and W. O. Nelson, Hypophyseal and Ovarian Hormones and Mammary Gland, "Ciba Foundation Colloquia on Endocrinology," Vol. 4, J. & A. Churchill, London, 1952, pp. 381, 395 and 402.

Mayer, G., and M. Klein, Physiologie de la lactation, *Ann. de la nutrit. et de l'aliment.*, **2**, 113, 1948; **3**, 667, 1949.

Meites, J., and C. W. Turner, Lactogenic Hormone, in C. W. Emmens, "Hormone Assay," Academic Press, New York, 1950.

Petersen, W. E., Lactation, *Physiol. Rev.*, **24**, 340, 1944.

Petersen, W. E., The Hormonal Control of Lactation, *Recent Progress in Hormone Res.*, **2**, 133, 1948.

Turner, C. W., in E. Allen, "Sex and Internal Secretion," 2d ed., Williams & Wilkins, Baltimore, 1939.

Turner, C. W., Mammogenic Hormone, in C. W. Emmens, "Hormone Assay," Academic Press, New York, 1950.

SECTION EIGHT

The Formation and Excretion of Urine

CHAPTER 63

The Kidney

THE KIDNEY IS undoubtedly the most important of all the organs that take part in the maintenance of the constancy of the internal environment, because, as Smith has remarked, "in the last analysis the composition of the *milieu intérieur* is determined not by what the body takes in, but by what is retained and what is excreted." The lung excretes CO_2 and water, and the skin water and a few salts. Through the intestine some of the heavy metals, the residue of food ingested, and the digestive secretions are excreted. The paramount importance of the kidney is due not only to the quantity, but also to the variety, of the substances eliminated in the urine.

Apart from the elimination of waste products of metabolism (nitrogen, sulfur, hormonal substances, etc.) and of toxic substances, the kidney by means of its excretory activity performs several other functions:

1. The regulation of the plasma volume and water content of the organism. When large amounts of water are ingested, the kidney responds by eliminating great quantities of dilute urine. When there is an abundant loss of water, due to profuse sweating, vomiting, diarrhea, etc., the kidney responds by excreting a smaller volume of greatly concentrated urine.
2. The regulation of the osmotic equilibrium and the maintenance of an optimal ionic balance in the plasma. The kidney performs this function by increasing or diminishing the excretion of such ions as sodium, potassium, calcium, chloride, phosphate, etc.
3. The regulation of the acid-base equilibrium of the organism. The kidney plays a part in this by eliminating a more or less alkaline or acid urine, in relation to the pH of the blood. Moreover, when nonvolatile acids are formed in the organism, the kidney produces NH_3 and thus spares the inorganic base in the blood.

The kidney also produces renin, an internal secretion, which plays a part in renal hypertension and seems to act in the normal regulation of blood pressure.

ANATOMY

In order to understand the processes by means of which urine is formed, it is necessary to describe briefly the structure of the kidneys.

The structural unit or nephron. The kidneys are made up of a great number of primary structural and functional units called nephrons (Fig. 288). These units are surrounded by connective tissue and are provided with an abundant blood supply. It has been calculated that there are about a million nephrons in each human kidney. Each nephron is made up of a malpighian body and a renal tube.

The malpighian body is situated in the cortex of the kidney and is formed by the blind end of a renal tube invaginated so as to contain a bundle of sinuous capillaries, which wind around each other; this is the glomerulus. The glomerulus is a capillary area placed between two arterioles, the afferent arteriole and the efferent arteriole, the former being larger than the latter. Both these arterioles have a well-developed muscular layer made up of circular smooth-muscle fibers. The visceral or internal layer of epithelium that covers the glomerulus is closely attached to the capillaries; the external or parietal layer forms a spherical capsule (Bowman's capsule) which contains the glomerulus. Between the layers there is a cleftlike space continuous with the lumen of the renal tube.

The renal tube arises in the malpighian body, to which it is joined by a short communicating

tube or "neck." It has three segments: (*a*) the proximal tube; (*b*) the thin segment; (*c*) the distal tube.

The proximal tube contains a pars convoluta situated in the renal cortex, near the glomerulus, and a pars recta, which goes toward the medulla and forms the first part of the descending branch of Henle's loop. The cells in the epithelium of the proximal tube have the shape of truncated pyramids, with a "brush border" on the free or luminal surface.

The thin segment is placed between the pars recta of the proximal tube and the pars recta of the distal tube, forming part of Henle's loop. This loop is that part of the renal tube which first goes toward the medulla (the descending branch) and then back to the cortex (the ascending branch). The thin segment occupies a variable position in the loop; it can form part of the descending branch, the ascending branch, or both. The epithelium of the thin segment is made up of flat cells with a clear cytoplasm and a nucleus that bulges into the lumen of the tube.

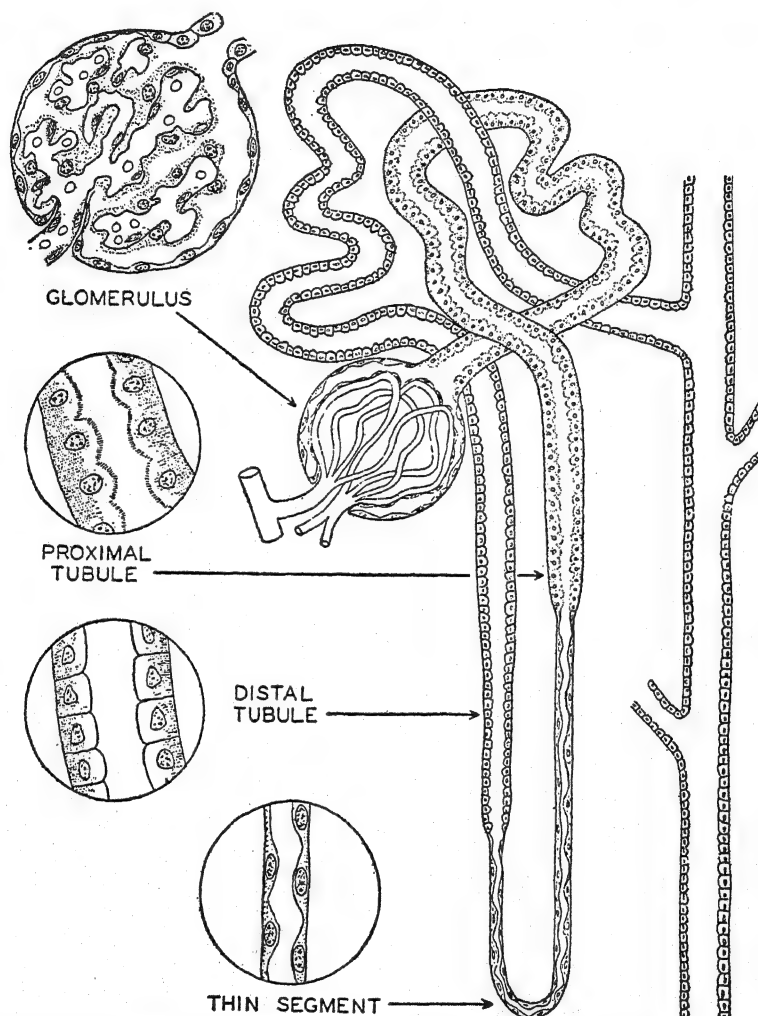



FIG. 288. Diagram showing the essential features of a typical nephron of the human kidney. (*Smith, H. W., "The Physiology of the Kidney," Oxford, New York, 1937.*)

The distal tube has a pars recta which follows the thin segment in the ascending branch of Henle's loops, and a pars convoluta continuous with the former and placed in the renal cortex. According to some observers the distal tube always comes into contact with the afferent arteriole of the glomerulus in which it begins,

and at this point the cells of the epithelium take on a columnar shape and the nuclei appear to be close to each other (macula densa). The epithelium of the distal tube is lower than that of the proximal tube; its cells project irregularly into the lumen of the tube, and they do not have a brush border. The distal tube is joined to the collecting tube by a short connecting segment.

The collecting tubes, so called because several nephrons open into each of them, descend, together with other collecting tubes, into the medulla and end in the cribriform area on the papillae of these pyramids. The fluid from the tubes flows into the minor calyces of the renal pelvis and thence through the ureter into the bladder. The connecting and collecting tubes are lined by an epithelium of cuboidal cells with clear cytoplasm and a nucleus that stains intensely.

The proximal tubes with a typical brush-border epithelium and the distal tubes are found in all animals. The thin segment is found only in the higher vertebrates (mammals and some of the birds). In the lower vertebrates this segment does not exist or is replaced by a short narrow tube lined by a ciliated epithelium called the intermediate segment (Fig. 289). The fact that the thin segment is found only in those animals which excrete urine of a higher osmotic pressure than that of the blood plasma has suggested the idea that in this segment most of the reabsorption of water takes place.

The differences in structure of the three segments of the nephron indicate functional differences, which have been demonstrated by physiological experiments (to be described further on) and which are substantiated by the distribution of the blood vessels of the kidney.

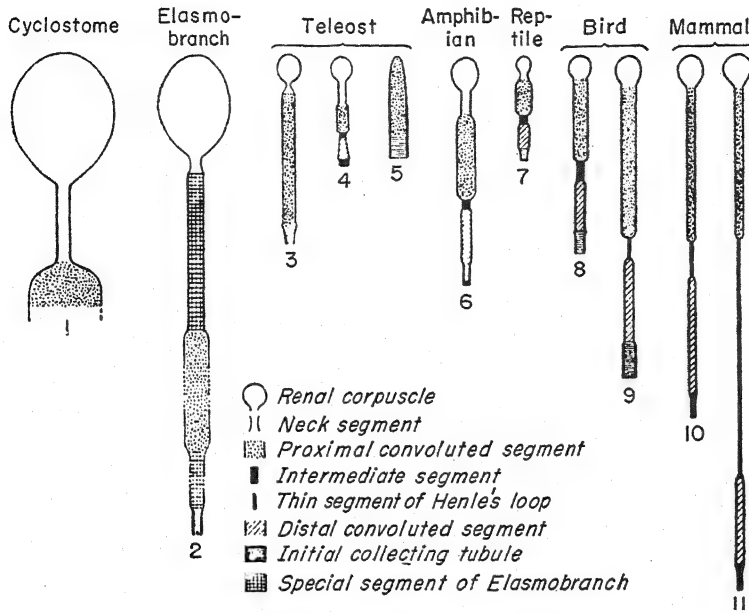



Fig. 289. Diagram of nephrons in different vertebrates. (*Marshall, E. K., Physiol. Rev., vol. 14, p. 133, 1934.*)

Blood vessels. The renal artery divides on arriving at the hilum of the kidney into ventral and dorsal branches, which again divide and subdivide, giving origin to the lobular arteries. These go to the base of Malpighi's pyramids where the arcuate arteries arise and run between the cortex and the medulla in a direction almost parallel to the surface of the kidney (Fig. 290). Each arcuate artery can be considered as a terminal vessel, because it does not anastomose with other arteries. Peripherally it gives out numerous ascending interlobular arteries from which the afferent arterioles of the glomeruli arise. These afferent arterioles have a well-developed muscular coat. In the neighborhood of the glomerulus the number of cells in the media and adventitia increase, forming a kind of cap on one side of the glomerulus (*Polkissen*). The cells of the *Polkissen* are free from granules and fibrils, and in certain species they have an epithelioid aspect. The term "juxtaglomerular apparatus" has

been given by Goormaghtigh to the formation constituted by the thickened vascular wall and the neighboring extravascular cells; he attributed an endocrine function to this apparatus. Others maintain that it modifies the lumen of the arteriole, thus regulating the blood flow through the glomerulus.

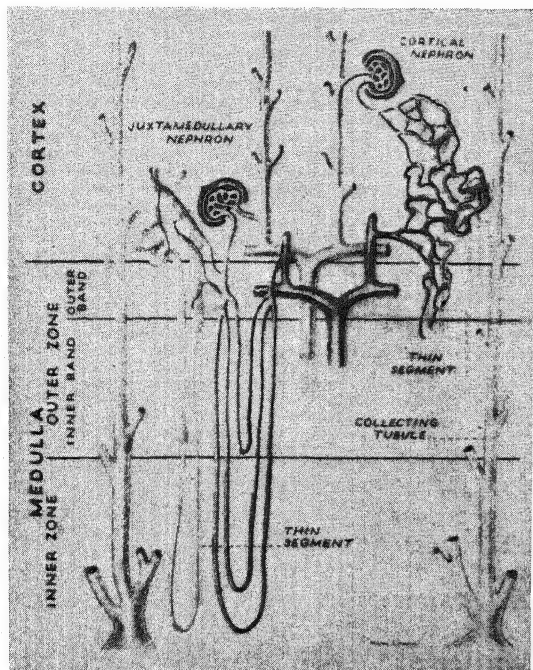



FIG. 290. Diagram showing significant differences between a cortical nephron (right) and one located in the juxtamedullary region (left). (*Smith, H. W., "The Kidney," Oxford, New York, 1951.*)

Trueta and his associates[1] have demonstrated that the ulterior course of the blood coming from the *cortical glomeruli* (situated in the external two-thirds of the cortex) differs from that taken by the blood from the *juxtamedullary g lomeruli* (situated in the internal third of the cortex). The efferent vessel of the former has a small caliber and soon divides into a capillary network surrounding the cortical tubes. The efferent vessel of the juxtamedullary glomeruli is of greater, or at least the same, caliber as the afferent vessel, and goes toward the medulla, where it divides into the *vasa recta*, large straight vessels which run parallel to the collecting tubes (Fig. 290).

The work of Trueta and his associates shows that these vessels can act as a short circuit in renal circulation. In certain physiologic and pathologic conditions, by means of a nervous mechanism the circulation in the cortex is considerably diminished, without any significant decrease in the renal blood flow. In these cases the vasa recta and the juxtamedullary glomeruli are found to be full of blood, while there is marked ischemia of the cortex. The juxtamedullary glomeruli and the arteriae rectae would therefore serve to short-circuit the blood when nervous and humoral factors constrict the afferent arterioles of the cortical glomeruli.

From the tubular capillaries the blood passes into the interlobular veins, which are originated in star-shaped groups on the renal surface. The interlobular veins, together with the ascending veins from the medulla, form the arcuate veins, which join to form the renal vein.

Between the tubes there is a complicated network of lymphatic capillaries, which flow into the deep lymphatics running next to the blood vessels and into the superficial lymphatics flowing from the surface of the kidney toward the hilum.

Innervation of the kidney. The kidney has a rich sympathetic innervation. Sympathetic fibers emerge from the spinal cord in the anterior roots of the fourth dorsal through the fourth lumbar segments. After passing through the prevertebral ganglia these fibers form a plexus around the renal artery. Some of the fibers in this plexus are preganglionic and end on nerve cells in the kidney. Sensory endings have been demonstrated in the adventitia of the blood vessels, and motor endings in the muscular coat. Nerve endings around the renal tubes and between the cells of the convoluted tubes have also been described.

Structural features. The following structural features of the kidney should be particularly noted: (*a*) the renal artery is wide and short, and soon branches out so that a high blood pressure is assured in the afferent arteriole; (*b*) the efferent arteriole of the cortical glomeruli is of a smaller diameter than the afferent arteriole, so that a high blood pressure is maintained in these glomeruli; (*c*) the renal tubes are supplied by a capillary network with blood at low pressure (this blood has already passed through the glomerulus); (*d*) the structural differences in the renal tube indicate the existence of functional differences in its several parts.

URINE[1]

Physical properties. The volume of urine formed daily by an adult man is from 1,000 to

[1] TRUETA, J., A. E. BARCLAY, P. DANIEL, K. J. FRANKLIN, and M. M. L. PRICHARD, *Lancet*, 2, 237, 1946; "Studies of the Renal Circulation," Blackwell, Oxford, 1947.

[1] More details are given in textbooks of biochemistry.

1,500 cc.; two-thirds to three-fourths is excreted during the day and the rest during the night. Children excrete three to four times more urine per kilogram of body weight than adults.

The volume of urine varies with the amount of fluid ingested and according to the amount of fluid excreted by other paths (skin, lung, intestine). Several other conditions influence the urinary volume:

1. Diet. A protein diet increases the amount of urea excreted and therefore the total volume of urine. In fasting less urine is excreted.
2. The temperature of the environment. A high temperature that provokes sweating diminishes the amount of urine formed.
3. Body posture. The urinary volume increases in the horizontal position.
4. Exercise (see Chap. 47, Changes in Renal Function).

The color of urine is golden yellow, the intensity of which increases with the concentration. It has an aromatic odor, which varies with the nature of food taken. The specific gravity at 15°C. is usually between 1.015 and 1.020, but can be 1.002 when large quantities of urine are excreted and 1.040 when the volume is small. The osmotic pressure is usually far above that of blood plasma. The freezing point of urine varies from −0.9 to −2.7°C., while that of plasma is −0.56°C.

The urine of man on a mixed diet is slightly acid; the pH is about 6. This is due to the end products of protein metabolism, which give several acids on oxidation (see page 459). The food of herbivora, on the contrary, gives many basic end products. The salts of tartaric, malic, and citric acids when oxidized set free fixed base; because of this the urine of herbivorous animals is alkaline. The urine of man on a vegetarian diet also becomes alkaline. The extreme values of urinary pH are 4.8 and 8.2.

Chemical composition. The chemical composition of urine is extremely variable. Its constituent substances can be classified as (*a*) inorganic; (*b*) organic.

Inorganic constituents. The inorganic constituents of urine are the acid ions (chloride, sulfate, and phosphate) and the basic ions (sodium, potassium, calcium, magnesium, and ammonia).

Urinary *chloride* has its source in the chloride of the food. The normal daily excretion of Cl is 6 to 10 gm. A diet rich in chloride increases this excretion, which diminishes or even ceases on a chloride-free diet.

Urinary *sulfates* (of Na, K, Ca, Mg) are produced almost exclusively by the oxidation of sulfur-containing amino acids, cystine, and methionine. A high-protein diet will provoke an increase in urinary sulfate, which is normally about 2 gm. daily (expressed as SO_3).

Urinary *phosphates* (of Na and K, and smaller amounts of Ca and Mg) originate partly in the inorganic phosphate of the food, but mostly in the metabolism of phosphorus-containing substances such as phosphoproteins, phospholipids, nucleoproteins, etc. The total phosphate (as P_2O_5) excreted daily in the urine varies from 1 to 5 gm.; this amount is dependent on the phosphorus-containing substances ingested and on the elimination of Ca and Mg phosphates in the feces. Within the pH range found in urine only monobasic and dibasic phosphates can be found. The relative amount of each is conditioned by the pH; as the pH rises so does the B_2HPO_4, and on the contrary, as acidity increases the BH_2PO_4 also increases. When the urine is alkaline, Ca and Mg phosphates are precipitated.

The amount of *sodium* and *potassium* excreted is dependent on the diet. Normally 2 gm. of K and 4 gm. of Na are eliminated daily. There are only small quantities of *calcium* (0.2 gm. per day) and *magnesium* (0.15 gm. per day).

Ammonia is a normal constituent of urine; the daily excretion is about 0.7 gm. The amount of NH_3 in the urine is conditioned by the excess of acid over base excreted. An increase in acid formation (high-protein diet, diabetic acidosis, etc.) results in an increase in urinary ammonia.

Organic constituents. The principal organic constituents of urine are urea, uric acid, creatinine, hippuric acid, urinary indican, urinary pigments, and lactic acid.

Urea, the principal nitrogenous substance in the urine, is an end product of protein metabolism. The greater part of the urea excreted in the urine is the result of the metabolism of ingested protein. The total daily excretion of urea is normally 25 to 35 gm. daily, but it varies with the protein intake.

Uric acid is a product of nucleoprotein metabolism. The exogenous uric acid varies with the diet, the endogenous uric acid comes from tissue destruction. In mammals (excepting man, the anthropoids, and the Dalmatian coach hound)

the end product of nucleoprotein metabolism is *allantoin*, resulting from the oxidation of uric acid. The daily excretion is 0.6 to 0.9 gm. In acid urine, uric acid has a tendency to precipitate.

Creatinine (1 to 2 gm. daily) is a normal urinary constituent (see Chap. 43, Protein Metabolism).

Hippuric acid is formed in the body by the combination of benzoic acid ingested with the food and glycine, which is split off from the proteins in the food or synthesized in the body. In the dog the kidney synthesizes hippuric acid, but in man and most other animals this operation is principally carried out by the liver. The daily excretion is 0.1 to 1 gm.; it increases when the diet contains much fruit and vegetables.

Urinary indican (potassium indoxylsulfate) comes from the indole produced from tryptophane in the course of intestinal putrefaction. The indole is absorbed, and in the liver it is oxidized to indoxyl, which combined with sulfuric acid gives urinary indican.

The principal *urinary pigment* is *urochrome*, the nature and significance of which are not well known. There is also *urobilinogen* derived from bilirubin by reduction in the intestine; it is oxidized into urobilin. The total amount of both these pigments is not more than 0.5 mg. daily; larger quantities are abnormal. In certain diseases bilirubin, hemoglobin, etc., can be found in the urine. Dyes such as methylene blue, prontosil, etc., injected or taken by mouth can be excreted in the urine, giving it their peculiar color.

THEORIES OF RENAL FUNCTION

Apart from a few simple experiments performed by Galen, no significant work on the mechanism of the formation of urine was done before the second half of the nineteenth century. Up to that time there were only surmises based on anatomical data. Galen, who was convinced of the truth of Aristotle's dictum that "nature does nothing in vain," when writing on the renal arteries stated, "If it is not to purify the blood they carry, tell me why nature has made them so large and why do they penetrate and ramify, the same as the veins, right into the depth of the renal cavities." Malpighi,[1] who discovered in the renal cortex small globular capsules containing a capillary knot which hung from the small arterioles "like apples on the branches of a tree," deduced that in these capsules (Malpighi's bodies) the urine was formed and then carried to the renal pelvis by the renal tubes. Malpighi's conclusions were not accepted by some of the writers at the beginning of the nineteenth century, but in 1842 Bowman[1] demonstrated the continuity of the capsule, which is now known by his name, with the renal tubes; he studied the epithelium of these tubes and remarked on its similarity to that of other glands. Based on these observations he formulated the hypothesis that the cells of the renal tubes eliminate from the blood the specific urinary constituents (urea, uric acid, etc.), while water (and salts) are filtered through the walls of the glomerular capillaries. This filtrate furnishes the water needed for the separation of urinous products from the epithelium of the tube.

Two years later Ludwig,[2] the famous German physiologist, in opposition to this vitalistic or physiological theory proposed one that was purely mechanical. He considered the glomerulus as a filtration apparatus. His theory of renal function is essentially the following: The difference between the pressure in the capillary and that in the cavity of Bowman's capsule causes an ultrafiltrate of plasma to pass through the capillary wall, *i.e.*, water and all the crystalloids, not only water and salts as Bowman believed. On passing through the renal tubes this ultrafiltrate is concentrated by reabsorption of water and certain salts into the blood of the peritubular capillaries. The low hydrostatic pressure within the capillaries, and the high oncotic pressure due to the loss of water as the blood passes through the glomerulus, assure the reabsorption of water by a process of "endosmosis."

In 1874 Heidenhain[3] modified Bowman's theory. He stated that the glomerulus secreted water and salts and that the tubular cells also secreted additional salts, metabolic waste products, and foreign substances. He and his pupils performed a great number of experiments to prove that the rate of urine formation depended on the rate of blood flow and not on the blood pressure in the glomerular capillaries; they also

[1] MALPIGHI, M., De viscerum structura, Bonn, 1666.

[1] BOWMAN, W., *Phil. Trans. Roy. Soc.*, **132**, 57, 1842.

[2] LUDWIG, C., in Wagner, "Handwörterbuch der Physiologie," vol. 2, 1844, p. 637.

[3] HEIDENHAIN, R., *Arch. f. mikr. Anat.*, **10**, 1, 1874; in Hermann's "Handbuch der Physiologie," **5**, 318, F. C. W. Vogel, Leipzig, 1883.

demonstrated that dyes injected into the blood stream were excreted into the lumen of the tubes by the cells of the tubular epithelium. For many years the philosophical bias of the investigators with regard to vitalism or mechanism had as much weight as the experimental evidence to incline them in favor of one or other of these theories.

Cushny[1] proposed, in 1917, what he then called the "modern view." He admitted that the first step in the formation of urine is a process of filtration through the glomerulus of a fluid, similar to blood plasma by its content of crystalloids and water, but free from proteins. This ultrafiltrate is concentrated as it passes down the renal tube, because a fluid of constant composition is reabsorbed. This fluid, a perfected Locke's solution, is returned to the blood, while in the tube remain and are concentrated the waste products to be eliminated in the urine. The reabsorption of this optimal fluid, according to Cushny, was due to the "vital activity of the tubular epithelium." He denied the capacity of the cells to excrete into the lumen of the tubes.

THE MECHANISM OF RENAL FUNCTION

Thanks to the important work of Richards, Rehberg, Marshall, Smith, Winton, Shannon, and others[2] the following facts have been demonstrated:

1. There is a process of filtration through the walls of the glomerulus, by which a fluid is formed similar in composition to the blood plasma, but without proteins and other large molecules; in other words, the fluid is an ultrafiltrate of plasma (glomerular filtration).
2. Along the renal tube, water and several substances (glucose, chloride, phosphate, urea, etc.) are reabsorbed, each one by a more or less independent process (tubular reabsorption).
3. The cells of the renal tubes can excrete certain substances from the blood into the tubular fluid (tubular excretion).
4. The cells of the renal tubes can produce certain substances and secrete them into the blood stream (renin) or into the tubular fluid (ammonia, hippuric acid).

[1] CUSHNY, A. R., "The Secretion of Urine," Longmans, New York, 1917.

[2] See the bibliography at the end of the chapter.

Each one of these functions of the kidney will now be considered in detail.

Glomerular Filtration

In the glomeruli, water and the crystalloids in the plasma are filtered through the glomerular

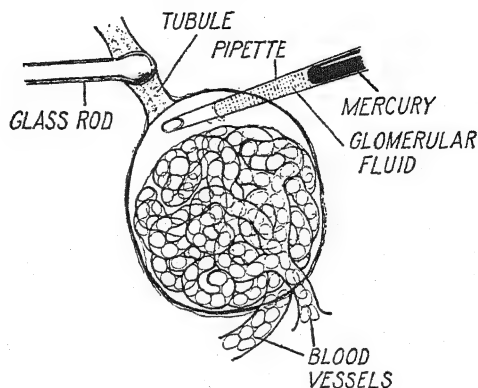



FIG. 291. Diagram of Richards' method for obtaining samples of glomerular filtrate.

membrane. Blood pressure is the only force acting in this process of simple ultrafiltration; the cells of Bowman's capsule and the capillary epithelium do not take an active part.

There is ample evidence of the truth of this statement. Ludwig and his pupils had demonstrated the relation between blood pressure and the amount of urine excreted. Heidenhain, who defended the secretory theory, maintained that variations in blood pressure were paralleled by variations in the rate of blood flow and that this was the important factor in determining the amount of urine secreted by the cells of the malpighian body. Later experiments by Richards and Plant[1] and others have shown that when a constant blood flow is maintained, the volume of urine diminishes when the blood pressure is lowered and vice versa.

Filtration takes place when the blood pressure in the glomerular capillaries is greater than the sum ot the osmotic pressure of the plasma proteins plus the pressure in Bowman's capsule:

$$Pb - (Po + Pc) = Pf$$

where Pb is the blood pressure, Po is the osmotic pressure of the plasma proteins, Pc is the hydrostatic pressure in Bowman's capsule, and Pf is the effective filtration pressure (Fig. 292).

[1] RICHARDS, A. N., and O. H. PLANT, *Am. J. Physiol.*, **59**, 144, 1922.

Many experiments have demonstrated the existence of this effective filtration pressure; *i.e.*, that the blood pressure exceeds the pressure opposed to the process of filtration and that the amount of filtrate is conditioned by the value of this effective pressure.

AFFERENT EFFERENT
BP 75 - OP 30
CAPS P 5
EFP 40

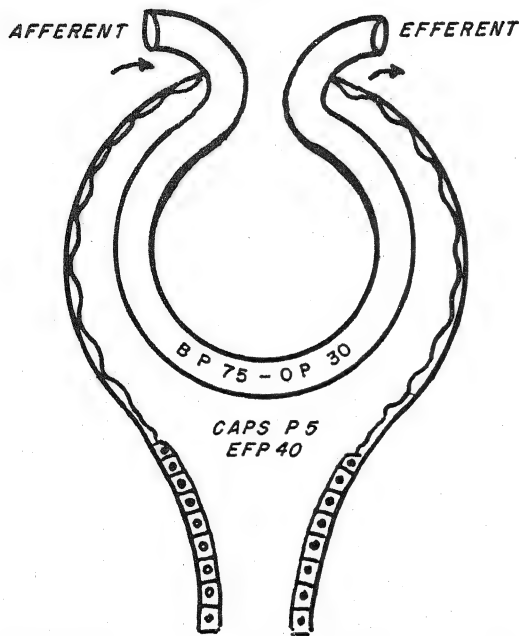



FIG 292. Diagram of factors in glomerular filtration.

Hayman[1] made direct measurements of the glomerular capillary pressure in the frog's kidney and found that it was 54 per cent of the mean aortic pressure. Winton[2] arrived at a similar figure (65 per cent) in the dog, using an indirect method. The osmotic pressure of the plasma proteins was first determined by Starling and others, who found values between 20 and 30 mm. Hg. The pressure in Bowman's capsule has been calculated to be from 5 to 10 mm. Hg. If the mean aortic pressure is 125 mm. Hg and the glomerular capillary pressure is 60 per cent of the aortic pressure, the effective filtration pressure is 40 mm. Hg.

$$Pb - (Po + Pc) = Pf$$
$$75 - (30 + 5) = 40$$

When the aortic pressure falls to 40–50 mm. Hg, and therefore *Pb* is only 24–30 mm. Hg, urine is not produced. The same result is obtained by artificially increasing the pressure in the ureter. On the other hand if the osmotic pressure of the plasma proteins is lowered by intravenous injection of saline solution, the urinary volume increases, and urine formation continues even when the blood pressure falls to 20 mm. Hg.

The pressure in the glomerular capillaries can change independently of the aortic pressure. This can occur when there is relaxation of the afferent arteriole; the resistance to the passage of blood into the glomerular capillaries is diminished, and thus the pressure in these capillaries rises. This takes place when the renal nerves are cut or when a drug that dilates the afferent arterioles is injected. Constriction of the efferent arteriole also increases the glomerular capillary pressure (adrenaline, splanchnic stimulation). Constriction of the afferent arteriole and dilatation of the efferent arteriole decrease the glomerular capillary pressure. Intravenous injection of Ringer's solution in the cat or the dog increases considerably the urinary flow without changing the oxygen consumption of the kidney. If the urine were formed by a process of secretion, there should be parallel changes in the urinary flow and the oxygen consumption, as the latter is conditioned by the activity of the cells.

The injection of proteins of different molecular weight has shown that the glomerular membranes are permeable to the smaller molecules. Thus, for example, hemoglobin (molecular weight, 68,000), egg albumen (35,000), Bence Jones protein (35,000), and gelatine (35,000) pass through the glomerular membranes, while serum albumin (70,000), serum globulin (170,000), casein (200,000), etc., do not. These data have served to indicate the size of the pores of the glomerular membrane, which is approximately 50A. When there is a pathologic disturbance that increases the permeability, serum albumin can be passed into the urine, but the larger serum globulins or fibrinogen very rarely filter through.

All these experiments are evidence in favor of the theory that in the glomerulus a process of simple filtration takes place. Richards and his collaborators have given a direct demonstration of this process of filtration. They used Chambers' technique of microdissection to obtain fluid from Bowman's capsule in the frog and necturus (Fig. 291). They then developed special methods to analyze the minute amounts of fluid collected

[1] HAYMAN, J. M., *Am. J. Physiol.*, **79**, 389, 1927.
[2] WINTON, F. R., *J. Physiol.*, **72**, 361, 1931.

and compared their composition with that of the blood plasma. They found that the fluid in Bowman's capsule in the frog and necturus is slightly alkaline, is free from protein, and contains glucose, chloride, and potassium, while the urine in these animals is acid and has neither glucose nor chloride. The capsular fluid has the same osmotic pressure, the same electrical conductivity, the same pH, and the same concentration of urea, glucose, inorganic phosphate, creatinine, and uric acid as the blood plasma. The concentration of chloride is somewhat higher, a fact that can be due to the Donnan equilibrium.

Until recently these experiments, which give a brilliant demonstration of glomerular filtration, could be held valid only for amphibians, and it was not legitimate to deduce from them that in mammals there is a similar process. Now, experiments in rats and guinea pigs have confirmed the fundamental facts found in amphibians. There is therefore sufficient evidence to maintain that through the glomerular capillaries a fluid free from proteins, containing the same crystalloids in the same concentration as in the blood plasma, is passed. This is exactly what occurs in the process of ultrafiltration of plasma through a semipermeable membrane.

Tubular Reabsorption

A comparison between the composition of the blood plasma, which is almost the same as that of the glomerular ultrafiltrate, and that of urine obtained from the bladder (Table 83) shows that some of the components of the plasma (or the glomerular ultrafiltrate) are not found in the urine (*e.g.*, glucose), others have been considerably concentrated, and others are in almost the same concentration in plasma and urine.

Table 83. Chemical Composition of Blood Plasma and of Urine

Constituent	*Blood plasma, mg. per 100 cc.*	*Urine, mg. per 100 cc.*	*Change in concentration in kidney*
Glucose	100		
Urea	30	2,000	60
Uric acid	2	50	25
Na	320	350	1
K	20	150	7
Ca	8	15	2
Mg	2.5	6	2
Cl	370	600	2
PO_4	9	270	30
SO_4	3	180	60

Source: Cushny, A. R., "The Secretion of Urine," Longmans, New York, 1917.

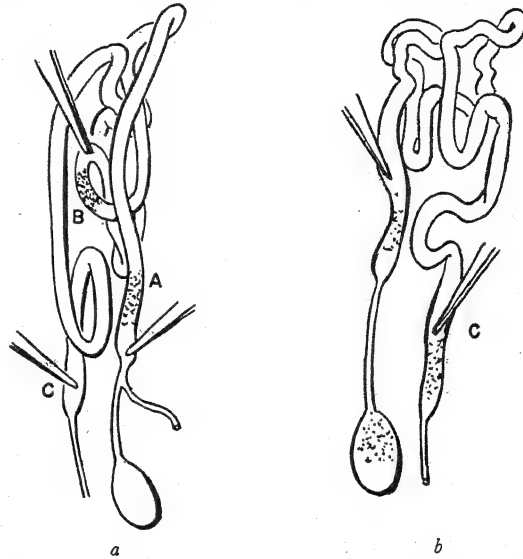



Fig. 293. *a*, collection of fluid from the upper (*A*), middle (*B*), and lower (*C*) portions of the renal tubes of necturus; *b*, perfusion of a single proximal tube of necturus. The shaded parts represent droplets of mercury placed to seal the tubes.

To explain the disappearance of glucose from the ultrafiltrate, it is necessary to admit that it has been reabsorbed as it passes down the renal tube. Changes observed in other components can be explained in either of two ways: (*a*) the tubes excrete concentrated solutions of some substances which are added to the glomerular ultrafiltrate; (*b*) the tubes, besides water, reabsorb selectively some of the substances in the glomerular filtrate, and the urine is the resulting residue of this filtrate.

Richards and his collaborators have given conclusive proof of reabsorption in the renal tubes of amphibians and mammals. Most of their experiments were made by collecting the fluid at different and well-located levels of the renal tube of the necturus, which has larger and less sinuous tubes than the frog (Fig. 293). To avoid contamination of the fluid in one part of the tube by that of other parts, the segment was blocked with a droplet of mercury placed in the lumen.

This method showed in the first place that the renal tubes of the necturus do not excrete water, as no fluid was collected from a segment separated from the glomerulus. The urine of this amphibian is extremely hypotonic; the fluid collected from the proximal tube and the intermediate segment had the same total concentration and chloride concentration as the blood plasma; on the other hand the fluid in the distal tube became more and more hypotonic, a fact that can be explained only by the reabsorption of osmotically active substances, especially of sodium chloride (Fig. 294). The urine becomes acid in the middle third of the distal tube. Glucose is reabsorbed almost exclusively in the proximal tube, and phlorhizin inhibits this process (Fig. 295). These experiments show that at least glucose and sodium chloride are reabsorbed in the tubes.

Walker and his collaborators[1] have obtained glomerular and tubular fluid from the kidneys of mammals (rat, guinea pig, and opossum) and have analyzed them by the microchemical methods employed by Richards, comparing the results with those obtained by the chemical analysis of blood plasma. The exact localization of the place punctured to collect the fluid is determined by subsequent dissection and isolation of the nephron. The fluid obtained from the glomerulus is usually free from protein; but even slight injuries can modify the permeability of the glomerular membrane and thus protein passes into the fluid. The concentration of glucose in the glomerular fluid is similar to that in the plasma (Fig. 296). It diminishes rapidly in the fluid obtained from successively lower levels of the proximal tube and disappears completely in the middle part of this segment. The reabsorption of glucose takes place in mammals, as in amphibians, in the first half of the proximal tube. Experiments in mammals injected with phlorhizin showed that this drug suppressed the reabsorption of glucose, as it did in amphibians; but in this case the concentration of glucose not only did not diminish, but increased progressively down the tube (Fig. 296). As there is no secretion of glucose, this increase can only be due to reabsorption of water. In experiments in rats and guinea pigs injected with 300 to 500 mg. per kg. of creatinine, an increasing concentration of this substance was found as the fluid progressed down the proximal tube. It has been calculated that, of 100 cc. filtered by the glomerulus, 80 cc. is reabsorbed in the proximal tube.

Two important facts have been observed by making a comparative study of the osmotic pressure and the chloride concentration in the tubular fluid with that of plasma:

1. The osmotic pressure of the tubular fluid does not change in the proximal tube, nor even in the first part of the distal tube. As a great amount of fluid is reabsorbed in the proximal tube, the reabsorbed fluid must have the same osmotic pressure as the plasma. The few results so far obtained by puncture of the distal tube seem to show that the reabsorption of water that results in the concentration of the urine takes place in the last part of the distal tube and not in Henle's loop as it was supposed up to now (Fig. 297).
2. Although an isosmotic fluid is reabsorbed in the proximal tube, chloride concentration increases progressively down this tube, but in the bladder the chloride concentration of the urine is lower than that of blood plasma. To explain these facts, it must be admitted (*a*) that there is an active and selective reabsorption in the proximal tube (the osmotic pressure does not change, although there is an increase in chloride concentration; therefore some other electrolyte, perhaps bicarbonate, is preferentially reabsorbed); (*b*) that there is reabsorption in the distal tube of a fluid that contains 1.4 the chloride concentration of the plasma (Fig. 297). Recent work[1] by Wesson, Anslow, and Smith has brought forward evidence in support of the hypothesis that a fluid isosmotic to blood plasma is reabsorbed in the proximal tube. Reabsorption of sodium in the proximal tube, however, does not depend on the reabsorption of water; therefore a fluid, the composition of which is constant, is not reabsorbed by this part of the renal tube, as was supposed by Cushny, and suggested by Pitts and Lotspeich[2] in the case of bicarbonate and chloride.

[1] Walker, A. M., and J. Oliver, *Am. J. Physiol.*, **134**, 562, 1941; Walker, A. M., P. A. Boot, J. Oliver, and M. C. MacDowell, *Am. J. Physiol.*, **134**, 580, 1941.

[1] Wesson, L. G., W. P. Anslow, and H. W. Smith, *Bull. New York Acad. Med.*, **24**, 586, 1948.

[2] Pitts, R. F., and W. D. Lotspeich, *Am. J. Physiol.*, **138**, 147, 1946; Lotspeich, W. D., R. C. Swan, and R. F. Pitts, *Am. J. Physiol.*, **148**, 445, 1947.

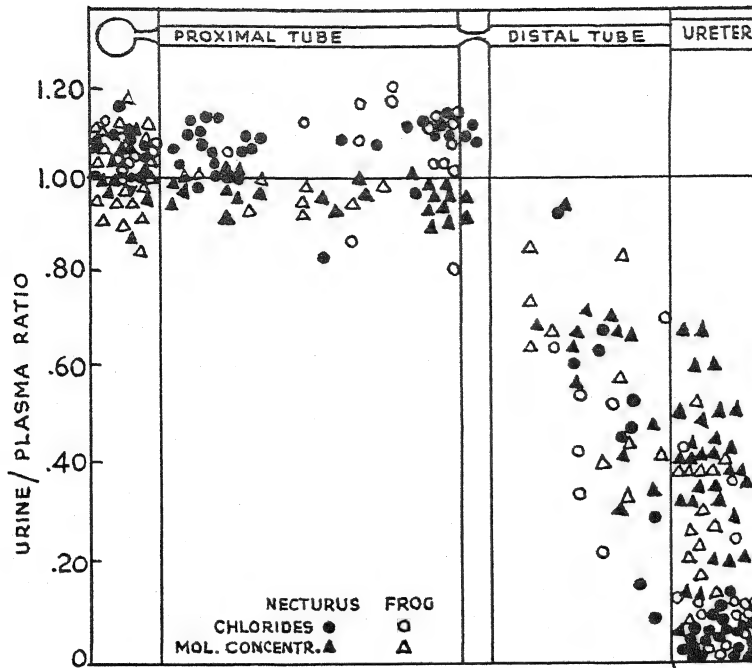



Fig. 294. Chloride and molecular concentrations in the glomerular filtrate are identical to those in plasma. In the proximal tube this identity is preserved; in the distal tube, chloride and necessarily sodium are reabsorbed. (*Adapted from Walker, A. M., C. L. Hudson, T. Findley, and A. N. Richards, Am. J. Physiol., vol. 118, p. 121, 1937.*)

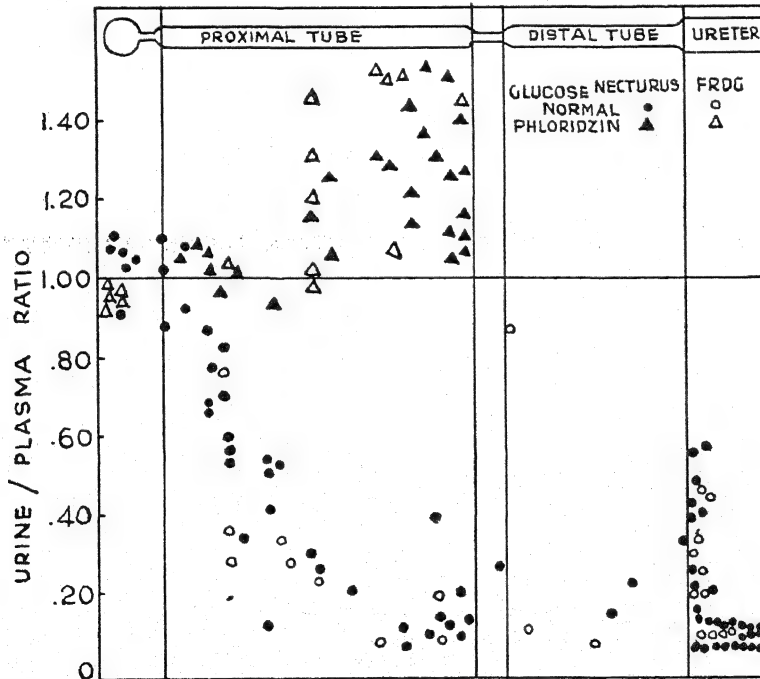



Fig. 295. The glucose in the glomerular filtrate is normally reabsorbed in the proximal tube of the necturus and frog. Adequate doses of phlorhizin block the tubular reabsorption of glucose, and as water is reabsorbed, glucose is concentrated in the proximal tube. (*Walker, A. M., and C. L. Hudson, Am. J. Physiol., vol. 118, p. 130, 1937.*)

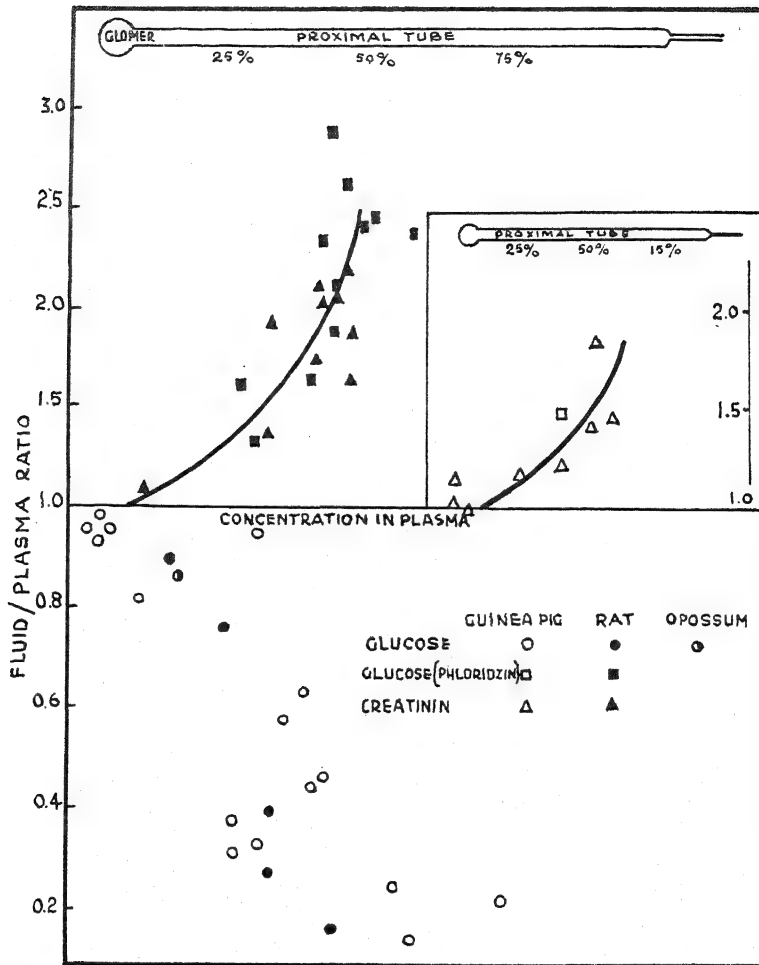



FIG. 296. Ratio of glucose (with and without previous phlorhizin injection) and creatinine concentration in plasma, in glomerular filtrate, and in the proximal tube. (*Walker, Bott, Oliver, and MacDowell, Am. J. Physiol., vol. 134, p. 562, 1941.*)

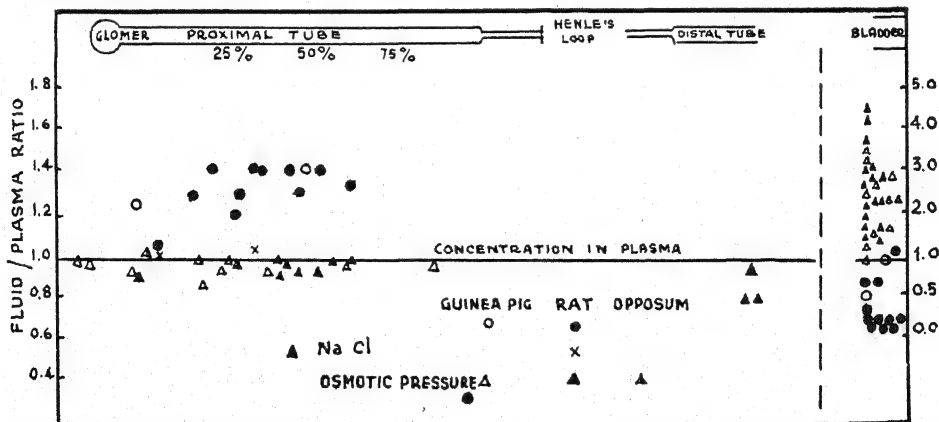



FIG. 297. Ratios of Cl and Na concentrations and osmotic pressure in blood plasma, glomerular filtrate, proximal and distal tubes, and in the bladder. On the left the figures correspond to glomerular filtrate and on the right to the bladder urine. (*From Walker, Bott, Oliver, and MacDowell, Am. J. Physiol., vol. 134, p. 562, 1941.*)

According to Wesson and his associates (Fig. 298), sodium, chloride, and bicarbonate are reabsorbed from the proximal tube by an active process which leaves a hypotonic fluid in the tube; water from this fluid is then reabsorbed passively. The functional significance of the thin segment would consist in the establishment of conditions favorable for the reabsorption of water, so that a fluid isosmotic to blood plasma would enter the distal tube. Approximately seven-eighths of the glomerular filtrate is reabsorbed in the proximal tube and the thin segment. Water and sodium are reabsorbed actively in the distal tube by processes which are independent of each other and which have a maximum that is not surpassed. The antidiuretic hormone promotes reabsorption of water in the distal tube.

FIG. 298. Passive and active reabsorption in the renal tubes.

TUBULAR EXCRETION

The expression "tubular excretion" will be used to signify the process by means of which the tubular cells take substances from the blood and discharge them into the lumen of the tubes. The word "secretion" would be equally appropriate, because the renal cells must perform work to transport substances from the blood to the urine. The word "excretion" is preferred because the process consists in the elimination of substances that have no longer any function in the organism.

Since the time of Heidenhain, many experiments have been made to prove the excretion of substances by the renal tube, which would in this way contribute to the formation of urine. Heidenhain injected indigo carmine intravenously under certain experimental conditions; the animals were killed and the kidneys examined microscopically; the pigment was found in the lumen of the tubes and in the cells of the proximal tube. This result would be obtained whether the dye were secreted by the cells, or filtered by the glomerulus and reabsorbed in the tubes.

The first conclusive demonstration that the tubes can excrete certain substances was given by Marshall and Grafflin[1] and by Edwards and Condorelli;[2] later Homer Smith and his collaborators gave further evidence of this property of the tubes. They studied the formation of urine in teleostean fishes, the kidneys of which have no glomeruli (aglomerular fishes) but consist essentially of tubes lined by an epithelium similar to that which lines the proximal tubes in amphibians and mammals. Water, creatine, creatinine, urea, uric acid, magnesium, potassium, sulfate, chloride, and several foreign substances are excreted. In these animals tubular excretion is the only mechanism for the formation of urine. On the other hand, they cannot excrete glucose and other carbohydrates, even after phlorhizin has been injected. The work in comparative anatomy and physiology of these investigators seems to prove that in the course of phylogenetic development the capacity

[1] MARSHALL, E. K., and A. L. GRAFFLIN, *Bull. Johns Hopkins Hosp.*, **43**, 205, 1928.

[2] EDWARDS, J. G., and L. CONDORELLI, *Am. J. Physiol.*, **86**, 383, 1928.

to excrete carbohydrate has never appeared in the renal tubes.

Chambers and Kempton[1] have also given a demonstration of tubular excretion. They cultivated *in vitro* the renal tubes of chick embryos. The extremities of the tubes are soon closed and the lumen is filled with fluid. Small quantities of phenol red (phenolsulfonphthalein) added to the culture medium are taken up by the tubular cells and excreted into the lumen of the tube, where the dye becomes considerably concentrated. On the other hand, phenol red injected into the tubes does not pass out into the culture medium.

To know if a substance in the urine is concentrated exclusively by the reabsorption of water or also by its excretion into the tubes, it is necessary to compare its concentration with that of a substance that is neither reabsorbed, nor excreted, nor synthesized by the tubes. Such a substance is glucose in animals intoxicated with phlorhizin. Let it be supposed that in an animal injected with phlorhizin the glucose concentration in blood is 100 mg. per cent and in urine 300 mg. per cent; glucose is therefore three times more concentrated in the urine than in the blood. If in the same experiment urea concentration in blood were 50 mg. per cent and in urine 400 mg. per cent, therefore eight times more concentrated, part of it must have been excreted by the tubes. If another substance (*e.g.*, inulin, creatinine) were concentrated three times, it would neither be reabsorbed nor excreted by the tubes; if yet another substance (*e.g.*, chloride) were found in the urine in the same concentration as in the blood it must have been reabsorbed in the tubes. There is satisfactory evidence that in the phlorhizinized animal glucose is neither reabsorbed nor excreted by the tubes; nevertheless these experiments are open to the objection that phlorhizin may disturb other functions of the tubes beside the reabsorption of glucose, and theretore in normal conditions the tubes perhaps behave in a different way. Further on it will be seen that there are other substances, among which is inulin, that are filtered through the glomerulus and are neither reabsorbed, nor excreted, nor synthesized by the tubes, nor do they disturb tubular functions.

In the phlorhizinized frog the ratio of the concentrations of glucose in urine and plasma (U/P) is 2.7:1; the concentration of urea in the urine is seven to eight times that of the plasma. This proves that in this amphibian urea is, at least in part, excreted by the tubes.

[1] Chambers, R., and R. T. Kempton, *J. Cell. & Comp. Physiol.*, **3**, 131, 1933.

The evidence for tubular excretion is therefore sufficient to admit that it exists, but the conclusions so far accepted cannot be applied to man in their entirety without further examination. Undoubtedly in mammals and in man the tubes can excrete certain substances, but "in respect to the relative importance of glomerular filtration, tubular reabsorption and tubular excretion, the human kidney must be examined on its own rights—and, in fact, the normal human kidney may be expected to differ considerably from the kidney that is altered by disease" (Homer Smith).

Active renal tubular transport. For a substance in solution to pass through an epithelium against the concentration gradient, the cells must perform work, the energy for which is obtained from intracellular metabolic processes. This type of secretion or absorption has been called "active transport."

The enzyme systems which take part in renal tubular active transport (secretion or absorption) are not yet well known. There is indirect evidence that glucose is reabsorbed in the proximal tube by successive processes of phosphorylation and dephosphorylation, activated by phosphorylase. Aerobic phosphorylation, according to Taggart,[1] Mudge, and others, plays a part in the excretion of para-aminohippurate, diodrast, and phenol red. Excretion of these substances is inhibited in the dog by 2,4-dinitrophenol, an inhibitor of aerobic oxidation and phosphorylation coupling in certain enzyme systems. Acetate and, to a lesser degree, pyruvate and lactate, which stimulate respiratory metabolism in slices of kidney, increase the rate of accumulation of para-aminohippurate in kidney slices; they also increase the maximum tubular excretion capacity for this substance.

Acetate seems to be an essential component of the enzyme system of tubular excretion. As it is found in very low concentrations in the cells and organic fluids, it is perhaps the limiting factor in tubular transport of the substances mentioned above.

[1] Taggart, J. V., Biochemical Aspects of Renal Tubular Transport, in "Renal Function," Josiah Macy, Jr., Foundation, New York, 1949.

Renal Clearance Tests

A method that has given important results in the study of renal function in man and other mammals is the determination of the renal clearance tests.

Ambard[1] made an experimental study of the relations between the supply and output of urea, and formulated his third law: "If the concentrations of urea in the blood and in the urine are both variable, the output of urea $(U \times V)$, in grams per 24 hr. is directly proportional to the square of the concentration of urea in the blood (P) in grams per liter, and inversely proportional to the square root of the concentration of urea in the urine (U) in grams per liter." This law can be expressed as follows:

$$(U \times V) = K \frac{P^2}{\sqrt{U}}$$

$$K = \frac{P}{\sqrt{(U \times V)\sqrt{U}}}$$

The constant K, called by Ambard the "ureosecretory constant," commonly known as "Ambard's constant," was used for many years to measure the functional capacity of the kidney. The value of K is 0.07 in a normal man of 70 kg. weight; in cases of renal insufficiency, K increases.

Later work showed that Ambard's laws were not strictly accurate. Austin, Stillman, and van

Fig. 299. "Standard" and "maximal" urea clearance and their relation to the filtration rate (*Smith, H. W., "The Physiology of the Kidney," Oxford University Press, New York, 1937.*)

The excretory activity of the kidney can be determined by measuring the total amount of a given substance eliminated in standard time. This is the renal output, easily calculated by multiplying the amount of urine secreted in 1 min. by the concentration of the substance in the urine. For example, if the rate of urinary secretion is 2.2 cc. of urine per minute (V), and if it contains 10 mg. of urea per cc. (U), the output of urea will be $2.2 \times 10 = 22$ mg./min.

The output is conditioned by the functional capacity of the kidney, but the most important factor is the concentration of the substance in the plasma (P), *i.e.*, the supply. This is easily understood if a substance that filters through the glomerulus is considered, and only its concentration in the glomerular filtrate is taken into account; thus if the ultrafiltrate is produced at a constant rate of 100 cc. per min. and the concentration of the substance in the plasma is 5 mg. per cent, the glomerular output will be 5 mg. per min.; if the concentration in the plasma is 10 mg. per cent, the output will be 10 mg. per min.

[1] Ambard, D., and A. Weill, *J. de physiol. et de path. gén.*, **14**, 753, 1912.

Slyke[1] observed that when the flow of urine was below 2 cc. per min. (the augmentation limit) the output of urea was proportional to the square root of the urine flow and not to the square root of the concentration of urea in urine, as Ambard had stated. When the urine flow was more than 2 cc. per min. (Fig. 299), the output of urea was directly proportional to the concentration of urea in the plasma (*i.e.*, to the supply), $U \times V = KP$, and the ratio of output to supply is constant $UV/P = K$. For example, if for a concentration of urea in plasma of 0.03 gm. per cent the output is 0.022 gm. per min., for a concentration of 0.06 gm. per cent, the output will be 0.044 gm. per min., etc.

What does this constant signify? Still using the figures in the example given, with an output of 0.022 gm. per min., urea concentration in urine 1 gm. per cent (U), urine flow 2.2 cc. per min. (V), and urea concentration in plasma 0.03 gm. per cent (P), the following result is obtained:

$$\frac{UV}{P} = K$$

$$\frac{1 \times 2.2}{0.03} = 73$$

where U is the concentration of urea in urine in grams per cent, P is the concentration of urea in plasma in grams per cent, and V is the urine flow in cubic centimeters per minute.

[1] AUSTIN, J. H., E. STILLMAN, and D. D. VAN SLYKE, *J. Biol. Chem.*, **46**, 91, 1921.

K is therefore the virtual quantity of plasma, in cubic centimeters, that would be completely deprived or cleared of urea by the activity of the kidney during 1 min.; in other words, the volume of plasma, in cubic centimeters, which contained the urea (or any other substance) excreted by the kidney during 1 min. The term "clearance" was first used in this sense by van Slyke and his collaborators[1] as an expression of renal function.

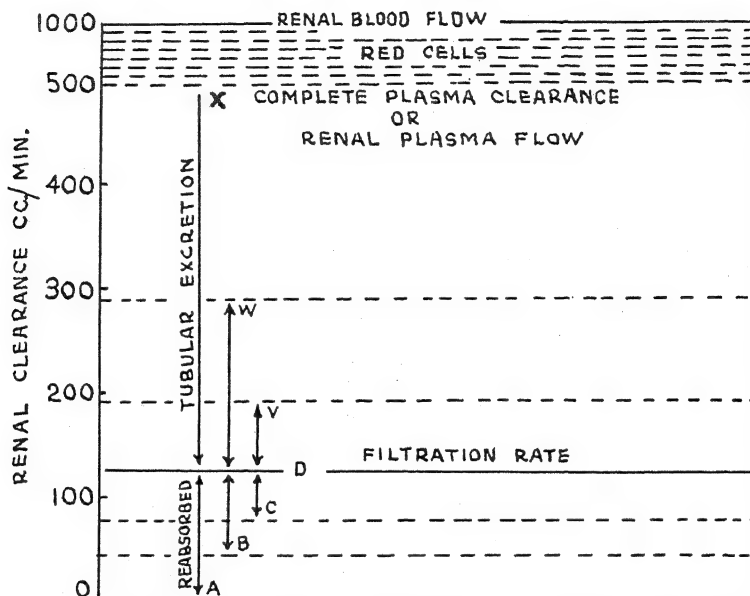



FIG. 300. Diagram illustrating the physiological concept of renal clearance. (*Smith, H. W., "The Physiology of the Kidney," Oxford University Press, New York, 1937.*)

The renal clearance of a substance is determined in the following way: The bladder is washed and completely evacuated, preferably by catheter. The urine excreted during a given time is collected and measured, and the concentration of the substance in the urine is determined. Several samples of blood are drawn at frequent intervals during the course of the experiment, and the concentration of the substance is determined. These values are plotted against time, and the concentration in blood at the middle of the period during which the urine is collected is accurately established. When the substance in question is not normally found in the blood, or exists in very small amounts, a certain quantity is injected or given

[1] MÖLLER, E., J. F. MCINTOSH, and D. D. VAN SLYKE, *J. Clin. Investigation*, **6**, 427, 1928.

by mouth. A constant plasma concentration can be kept up by continuous intravenous infusion of the substance.

The renal clearance of any substance can be determined. Experiments in man and in animals have demonstrated that this is an accurate and sensitive method of determining the conditions of renal function. Moreover by comparing the clearance of different substances several facts have been discovered, which are of importance in the understanding of renal function.

Not all substances have equal clearances. For example, urea clearance is 75 (*i.e.*, 75 cc. of plasma contain the urea eliminated in 1 min. of renal activity); inulin clearance is from 120 to 130; diodrast[1] clearance is 500 to 600. As will be shown later, the differences in clearance are due to the different ways in which the kidneys eliminate these substances. There are three possible ways in which this elimination can be performed: (*a*) glomerular ultrafiltration; (*b*) glomerular ultrafiltration and tubular excretion; (*c*) glomerular ultrafiltration and tubular reabsorption.

A substance in the plasma that filters through the glomerulus but is completely reabsorbed in the tubes is not found in the urine, *e.g.*, glucose in normal conditions. In this case the clearance will be zero (Fig. 300*A*). If tubular reabsorption is incomplete (urea), the substance will appear in the urine and its clearance will increase as reabsorption diminishes (Fig. 300*B* and *C*). If there is no reabsorption (inulin) the clearance will be equal to the volume of the glomerular filtrate in 1 min., *i.e.*, the rate of glomerular filtration (Fig. 300*D*).

This last case can be better understood by taking an example. Suppose the glomeruli filter at the rate of 125 cc. per min., and a substance that passes through the glomerulus and is not reabsorbed by the tubes is found in the plasma in a concentration of 100 mg. per cent. In 1 min., 125 mg. of the substance will pass into the tubes, where water will be reabsorbed. The volume of urine excreted per minute will be much smaller than that of the glomerular filtrate, but the substance will be eliminated in the urine at the rate of 125 mg. per min. If the rate of excretion of urine is 10 cc. per min., the concentration of the substance will be 1,250 mg. per cent; if it is 2 cc. per min., it will be 6,250 mg. per cent; but in any case $UV = 12{,}500$, and as $P = 100$, $UV/P = 12{,}500/100 = 125$ cc. In this case the plasma volume that contained the amount of substance eliminated, *i.e.*, the clearance, *equals the volume of glomerular filtrate* (Fig. 301).

In the case of a substance that not only filters through the glomerulus but also is excreted by the tubes, the clearance will increase with the tubular excretion. Maximum or complete clearance is reached when all the blood that passes through the kidney has been cleared of the substance (Fig. 300*X*), thus no longer containing any when it leaves the kidney by the renal vein. In this case the *clearance equals the volume of plasma that flows through the kidney in 1 min.* (renal plasma flow). Actually the clearance for most substances is well below the plasma flow.

The clearance of any substance considered by itself would therefore have little value, as different excretory operations take part in the elimination of the substance. On the other hand if a substance has a clearance equivalent to the glomerular filtration rate, it can serve by comparison with the clearances of other substances to determine what factors play a part in the elimination of these. Glucose clearance in the phlorhizinized animal is equivalent to the glomerular filtration volume, but this method has drawbacks that have already been discussed.

Glomerular Filtration Rate

According to Smith,[1] a substance suitable for measuring the glomerular filtration rate must fulfill certain specifications:

1. It must be completely filterable at the glomerulus; *i.e.*, its molecular size must be such that it can pass through the glomerular membranes, and it must not combine with the unfilterable plasma proteins.
2. It must not be reabsorbed, excreted, or synthesized by the tubes.
3. It must be physiologically inert, so as not to produce any damage to the organism in general and to the kidney in particular.
4. It must be of a chemical nature such that its concentration in plasma and urine can be easily and accurately determined.

[1] 3:5 diiodo-4-pyridon-N-acetic diethanolamine.

[1] Smith, H. W., "The Physiology of the Kidney," Oxford, New York, 1937.

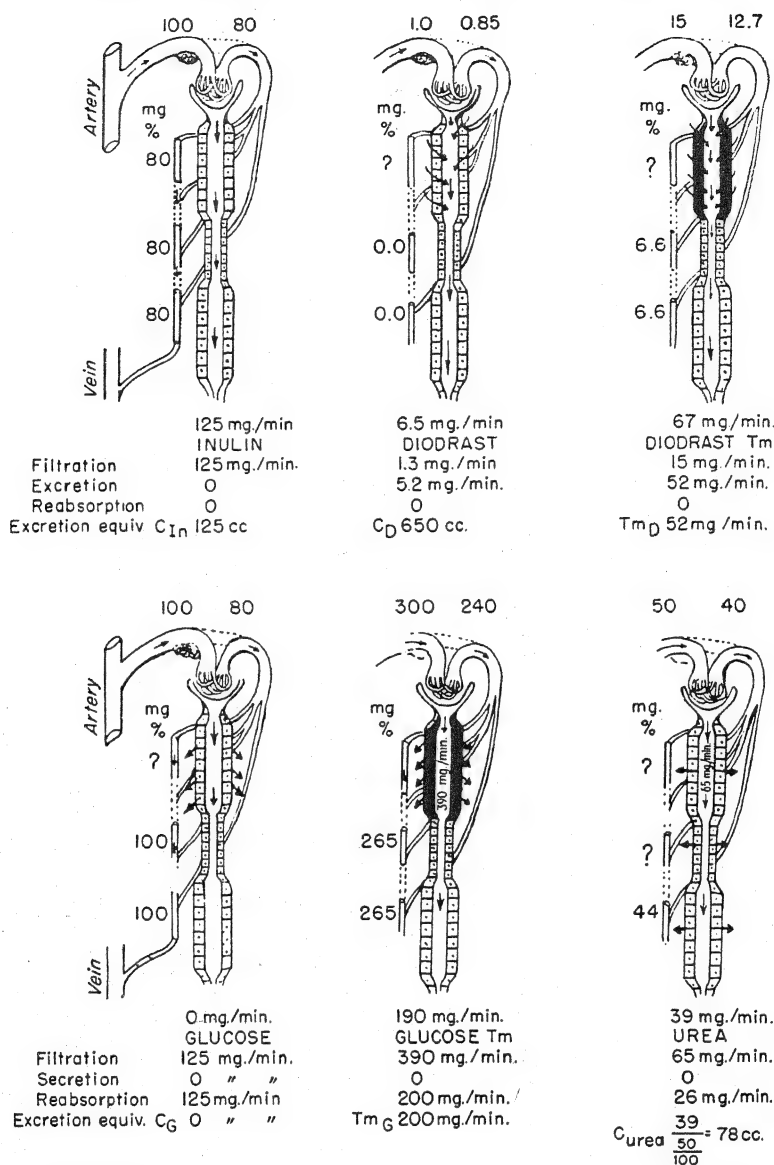



FIG. 301. Diagrams illustrating mechanisms of inulin, diodrast, glucose, and urea excretion. One nephron is diagrammatically represented, and the concentration, in milligrams per cent, of the different substances in plasma, the afferent and efferent arterioles, and the renal vein. The arrows indicate the direction in which the substance goes. The blackened parts represent cells functioning at a maximum. The figures given are only approximately those observed. C_{In}, C_D, and C_G represent inulin, diodrast, and glucose clearances, respectively; Tm_D and Tm_G, diodrast and glucose tubular mass. (*Adapted from Corcoran, A. C., West. J. Surg., vol. 51, p. 622, 1942.*)

Rehberg,[1] who was the first to measure the glomerular filtration rate using such a substance, believed that creatinine fulfilled the requisite conditions. Later work showed that in man creatinine is in part excreted by the tubes; therefore its clearance is greater than the glomerular filtrate. In the dog, rabbit, rat, sheep, and seal it is eliminated exclusively through the glomerulus and is not reabsorbed by the tubes.

Smith and his collaborators[1] and Richards

[1] REHBERG, P. B., *Biochem. J.*, **20**, 447, 1926.

[1] JOLIFFE, N., J. A. SHANNON, and H. W. SMITH, *Am. J. Physiol.*, **100**, 301, 1932; SMITH, H. W., The Excretion of

and his collaborators,[1] independently of each other, found that inulin is one of the substances that most satisfactorily corresponds to the specifications required for measuring the glomerular filtration rate.

FIG. 302. Diagrams illustrating clearance of several substances.

Inulin is a polysaccharide similar to starch, which on hydrolysis gives fructose. It is almost insoluble in cold water but is dissolved by hot water, which on cooling gives a supersaturated solution. It cannot be given by mouth, as creatinine can, but must be injected, because the digestive enzymes destroy it. The molecular weight of inulin is 5,000, the equivalent of 32 molecules of hexose. It is inert, and the blood has no enzymes that destroy it. The glomerular membranes do not retain inulin, but because of the size and elongated shape of its molecules it does not diffuse through the tubular walls. Besides creatinine and inulin, manitol, sodium thiosulfate, and several other substances have been used to measure the glomerular filtration rate.

Inulin clearance in a man of 1.73 sq. m. body surface is around 120. This means that the glomerular filtration rate is 120 cc. per min.; therefore 170 liters per day. The volume of urine excreted per day is 1.5 liters. Therefore 168.5 liters, 99 per cent of the glomerular filtrate, is reabsorbed by the renal tubes.

In basal conditions the glomerular filtration rate is fairly constant. The variations in the volume of urine excreted are therefore due to variations in water tubular reabsorption, which, as shown later, is regulated by the hypophysis.

By comparing the clearance of a substance x with that of inulin, many useful facts on the mechanism of excretion of the substance in question can be obtained. Thus if the ratio of x clearance to inulin clearance is less than 1:1, x is reabsorbed by the tubes. If this ratio equals 1:1, x is simply filtered at the glomerulus. If the ratio is greater than 1:1, x is not only filtered through the glomerulus but also excreted by the tubes.

Urea clearance, when the rate of urine excretion is above 2 cc. per min. (maximum urea clearance), is 75 cc. The ratio of urea clearance to inulin clearance is $75/120 = 0.6$, showing that urea is filtered through the glomerulus and then reabsorbed in part in the tubes, apparently by simple diffusion (Figs. 301, 302, and 303).

Renal Plasma Flow

Diodrast is an organic iodine compound used for x-ray visualization of the urinary tract. Its clearance is considerably greater than that of inulin; therefore besides filtering through the glomerulus it is excreted by the tubes. When the plasma concentration of diodrast (calculated from the plasma concentration of iodine) is high, its clearance equals UV/P. When its concentration in plasma decreases (*i.e.*, when the value of P decreases), the clearance increases up to a maximum reached when the plasma iodine is about 0.8 to 2 mg. per cent. When the plasma

Non-metabolized Sugars in the Dogfish, the Dog and Man, in Berglund and Medes, "The Kidney in Health and Disease," Lea & Febiger, Philadelphia, 1935.

[1] RICHARDS, A. N., P. B. WESTFALL, and P. A. BOTT, *Proc. Soc. Exper. Biol. & Med.*, **32**, 73, 1934.

concentration is below 2 mg. per cent there is complete plasma clearance of diodrast, *i.e.*, there is none in the blood of the renal vein (Fig. 301). The diodrast clearance (C_D) in this case equals the total plasma flow through the active renal tissue. The average flow in man is 669 cc. per min.[1] The blood flow through the kidney can be calculated by determining the plasma-cell ratio in the hematocrit. On an average, the blood flow is 1,115 cc. per min. If these determinations are accurate, approximately 25 per cent of the cardiac output (±5 liters per min.) goes to the kidneys, although these organs make up only 0.42 per cent of the body weight.

The renal circulation has a considerable degree of autonomy. The blood flow through the kidney is not greatly influenced by changes in blood pressure, even after the renal nerves have been cut. There are only three known ways of increasing the renal blood flow: (*a*) unilateral nephrectomy in the dog and in man causes an increase of 50 to 70 per cent in the renal blood flow of the remaining kidney; (*b*) in the unilaterally nephrectomized dog a meat diet increases the renal blood flow (*c*) certain pyretic agents (antityphoid vaccine, pyrogenic inulin) increase the renal blood flow in man.[2]

[1] GOLDRING, W., H. CHASIS, H. A. RANGES, and H. W. SMITH, *J. Clin. Investigation,* **19**, 739, 1940.
[2] SMITH, H. W., *Harvey Lect.,* **35**, 166, 1939–1940.

THE TUBULAR EXCRETORY MASS

When the diodrast concentration in plasma increases, its clearance diminishes (Figs. 302 and 303). At a plasma-iodine concentration of 30 mg. per cent or more, the tubes are excreting at full rate, and tubular excretion of diodrast reaches a maximum that cannot be surpassed. The amount of diodrast excreted ($U \times V$) in milligrams of iodine per minute, minus the amount eliminated by glomerular filtration (which can be calculated from the inulin clearance, the concentration of diodrast in blood, and a factor dependent on the concentration of albumin in plasma, which is 0.72 ± 0.025), gives the maximum amount excreted by the renal tubes. This amount is expressed in milligrams of iodine excreted per minute (on an average, 51.6 mg.) and is called the tubular excretory mass (T_m) (Fig. 301).

Smith and his collaborators[1] have found other substances that behave like diodrast and can therefore be used to measure the renal plasma flow and the tubular excretory mass. The one most frequently used is the sodium salt of para-aminohippuric acid.

Not only can the tubular excretory mass be measured, the maximal rate of reabsorption can also be determined. For example, there is satis-

[1] SMITH, H. W., N. FINKELSTEIN, L. ALIMINOSA, B. CRAWFORD, and M. GRABER, *J. Clin. Investigation,* **24**, 388, 1945.

FIG. 303. Variations in diodrast and glucose clearances when the plasma concentration of these substances increases.

factory proof that when the concentration of glucose in blood is sufficiently high, the rate of reabsorption of glucose by the tubes reaches a maximum that cannot be surpassed[1] (Figs. 302 and 303). The maximal reabsorption rate of glucose by the proximal tubes can thus be measured (Fig. 301). This constant maximal reabsorption rate of glucose is determined by the clearance methods and is known as the glucose Tm. It is interesting to note that the simultaneous determination of the Tm of glucose and that of para-aminohippuric acid shows a reciprocal depression of both functions. This seems to occur with many other substances, and the fact has been used to retard the excretion of certain drugs by the renal tubes (*e.g.*, penicillin). The depression is attributed to competition for the available cellular energy, more than to competition for a substance in the cell which takes part in both the process of tubular reabsorption and that of excretion.

Table 84. Mean Values of Renal Functional Measurements in Normal Subjects, by Sex

Measurement	*Males*		*Females*	
	Number of subjects	*Mean value*	*Number of subjects*	*Mean value*
Rate of glomerular filtration (inulin clearance, C_{In}), cc. per min	67	131 ± 21.5	21	117 ± 15.6
Effective renal plasma flow (diodrast clearance, C_D), cc. per min	61	697 ± 135.9	17	594 ± 102.4
Maximal tubular excretory capacity (diodrast Tm, Tm_D), mg. iodine per min	40	51.8 ± 8.73	14	42.6 ± 9.46
Maximal tubular reabsorptive capacity (glucose Tm, Tm_G), mg. glucose per min	24	375 ± 79.7	11	303 ± 55.3
Filtration fraction, C_{In}/C_D	61	0.19 ± 0.02	17	0.20 ± 0.03
$\frac{\text{Effective renal plasma flow}}{\text{Maximal tubular excretory capacity}} \left(\frac{C_D}{Tm_D}\right)$	34	14.0 ± 2.16	14	14.2 ± 2.36

Source: GOLDRING, W., and H. CHASIS, "Hypertension and Hypertensive Disease," Commonwealth Fund, Division of Publication, New York, 1944.

Note: All values are corrected to 1.73 sq. m.

Renal indexes. The ratio of diodrast clearance to diodrast tubular mass, C_D/Tm_D, expresses the virtual quantity of plasma completely cleared of diodrast by each unit of excretory tissue taking part in the operation. This ratio is fairly constant in normal subjects, and it gives a more accurate picture than the ratio of C_D to the body surface, as there is no constant ratio of kidney weight to body surface, and the amount of active tissue varies. When C_D/Tm_D is below normal there is renal ischemia.

The ratio of inulin clearance to diodrast clearance, C_{In}/C_D, known as the "filtration fraction," indicates the relative amount of plasma filtered through the glomerulus and gives information on the tone of the glomerular efferent arterioles. When the glomerular capillary pressure increases because of constriction of the efferent arterioles, C_{In}/C_D increases, and vice versa.

An increase in renal blood flow caused by a pyretic substance is accompanied by a decrease in the filtration fraction; it is therefore due in great part to dilatation of the efferent arterioles. On the contrary, orthostatic vasoconstriction and that produced by adrenaline injection or of psychic origin are accompanied by an increase in the tone of the efferent arteriole and therefore an increase in the filtration fraction.

The normal values corresponding to these different ratios are given in Table 84.[1]

CLINICAL TESTS OF RENAL FUNCTION

Several clinical tests, based on renal physiology, are used to obtain information as to the normal or abnormal condition of renal function. The more commonly used tests measure the capacity of the kidney to concentrate the urine.

Concentration and dilution. Volhard first proposed this simple and informative test. The

[1] SHANNON, J. A., and S. FISCHER, *Am. J. Physiol.*, **112**, 765, 1938.

[1] GOLDRING, CHASIS, RANGES, and SMITH, *loc. cit.*

original procedure has been modified by several observers, and it is now as follows: After emptying his bladder, the patient drinks 1,500 cc. of water or weak tea in the course of 30 to 45 min. He is kept in bed and is made to empty his bladder every ½ hr. The volume and specific gravity of each fraction of urine are measured. A normal subject eliminates the 1,500 cc. of water in 2 to 3 hr., or 4 hr. at the latest. The specific gravity can drop to 1.002. After 4 hr. the patient is fed on a dry diet. The specific gravity of the urine increases rapidly and reaches 1.030 in the course of the same day or night. One of the first signs of renal insufficiency is the loss of power to concentrate the urine, a function of the renal tubes. In cases of glomerular lesions the kidney cannot secrete large quantities of dilute urine.

Forced tubular reabsorption.[1] This test is performed as follows: The patient drinks 1 liter of water in the course of 10 min., and the volume of urine passed in the second and third half hours following the ingestion is measured. On the next day the same amount of water is taken after an intramuscular injection of 5 vasopressor units of pitressin (0.5 cc.). A normal subject eliminates about 200 cc. of urine in each half-hour period of the first day (without pitressin) and about 20 cc. on the second day (with pitressin). The reabsorption coefficient $\left(\frac{\text{diuresis without pitressin}}{\text{diuresis with pitressin}}\right)$ is greater than 5 and usually 10 to 12 in normal subjects. Low values are observed in cases where a certain number of renal tubes are damaged or are functionally incompetent.

Urea clearance.[2] Urea clearance is frequently used as a clinical test of renal function. It must be kept in mind that the normal figure varies with the quantity of urine excreted. When diuresis exceeds 2 cc. per min. (augmentation limit) the clearance does not increase with the rate of urine formation but remains fairly constant. This maximal urea clearance is determined by the formula $(U \times V)/P$, and is 75 cc. per min. in normal subjects. When diuresis is below 2 cc. per min., urea clearance is less and is dependent on the rate of urine formation. This standard urea clearance responds to the formula $(U \times \sqrt{V})/P$ and is 54 cc. in normal subjects. Urea clearance is measured, as was explained above, applying one or other formula according to whether the rate of urine excretion is above or below 2 cc. per minute. Usually the results are given as a percentage of the normal figure; the value found is multiplied by $100/75$ or 1.33 for maximal urea clearance, and by $100/54$ or 1.85 for standard urea clearance. The determination of urea clearance is the most valuable of all the clinical tests so far in use for the early diagnosis of renal insufficiency.

[1] Pasqualini, R. Q., and E. Etala, *Rev. Soc. argent. de biol.*, **15**, 161, 1940.

[2] Möller, E., J. F. McIntosh, and D. D. van Slyke, *J. Clin. Investigation*, **19**, 739, 1940.

FUNCTIONS OF THE KIDNEY

The Excretion of Water

The rate of urine excretion depends on the one hand on the rate of glomerular filtration and on the other on the rate of tubular reabsorption. The rate of glomerular filtration is conditioned by several factors. The most important are the following: (*a*) the permeability and filtration area of the glomerular membranes; (*b*) glomerular capillary pressure, dependent on arterial blood pressure and the tone of the afferent and efferent glomerular arterioles; (*c*) osmotic pressure of plasma proteins; (*d*) the number of active glomeruli. Recent work has shown that in man and other mammals all the nephrons seem to be permanently active; they do not function alternately as in the amphibian kidney.

The existence of these factors that can influence glomerular filtration rate (measured by inulin clearance) does not prevent it from remaining relatively constant in the dog and in man, even in conditions that cause widely diverging urinary excretion rates. On the other hand the glomerular filtration rate varies as the diuresis—in other words, with the degree of hydration—in the newborn child,[1] the rat,[2] and the rabbit.[3]

The renal tubes reabsorb always at least 100 cc. of the 120 cc. passing through the glomeruli in 1 min., as even in the most severe cases of diabetes insipidus the urine excreted does not exceed 20 cc. per min. This "obligatory" reabsorption of 80 to 87 per cent of the total glomerular filtrate takes place mostly in the proximal tube, along which the urine remains

[1] McCance, R. A., and W. F. Young, *J. Physiol.*, **99**, 265, 1941.

[2] Braun-Menéndez, E., and H. Chiodi, *Rev. Soc. argent. de biol.*, **22**, 314, 1946.

[3] Dicker, S. E., and H. Heller, *J. Physiol.*, **103**, 449, 1945.

isosmotic with the blood plasma. At this level reabsorption of water is *passive;* it diffuses through the epithelium together with sodium, chloride, bicarbonate, glucose, etc., and the urine is not concentrated. The partial reabsorption of the remaining 13 to 20 per cent is "facultative," and takes place principally in the distal tube, although the collecting tubes may play a part in this process. This second fraction of water is reabsorbed against the osmotic pressure of the urine (active reabsorption), which is thus concentrated. The energy necessary to produce a hypertonic fluid is probably provided by metabolic operations in the cells of the renal tubes.

Facultative reabsorption of water (as distinct from the obligatory reabsorption of five-sixths of the glomerular filtrate) is regulated by the antidiuretic hormone of the posterior lobe of the hypophysis. This hormone acts on the cells of the distal tubes, increasing the rate of water reabsorption, and is thus responsible for the urine having a higher osmotic pressure than plasma. The term "antidiuretic hormone" is an appropriate one, as the diuretic response normally following the ingestion of water is suppressed by this hormone. On the other hand it does not diminish diuresis provoked by the ingestion of salt or urea; therefore its effect is to promote the reabsorption of water not needed to excrete dissolved substances. Marshall maintains that the antidiuretic hormone has no effect in those animals which have no thin segment.[1] The antidiuretic hormone acts directly on the kidney, without the mediation of the nervous system; its effect is not altered by cutting the renal nerves, and it is observed in transplanted kidneys. When the hormone is no longer secreted, as occurs after cutting the hypophyseal stalk or removing the neurohypophysis and in diabetes insipidus, there is polyuria and hypotonic urine is excreted; five-sixths of the glomerular filtrate is still reabsorbed, but facultative reabsorption of water ceases. There seems to be a continuous secretion of this hormone (1 to 5 milliunits per hour in the dog) in normal conditions. The secretion is dependent on the integrity of the central nervous system[2] (see "Functions of the neurohypophysis," Chap. 52).

[1] BURGESS, W. W., A. M. HARVEY, and E. K. MARSHALL, *J. Pharmacol. & Exper. Therap.*, **49**, 237, 1933.

[2] O'CONNOR, W. J., and E. B. VERNEY, *Quart. J. Exper. Physiol.*, **31**, 393, 1942.

Functions of the Renal Nerves

The wealth of the innervation of the kidney suggests that it must play some part, perhaps an important one, in the functions of the renal tubes and in glomerular filtration. Nevertheless it has been repeatedly demonstrated that a single completely denervated kidney or a transplanted kidney can keep an animal alive and without any sign of renal insufficiency. This single denervated kidney responds normally to all internal and external factors that modify renal function.

Claude Bernard observed that denervation of the kidney provokes polyuria, a fact that has been repeatedly confirmed. More recent well-controlled experiments by Verney and his collaborators[1] in unanesthetized normal animals have shown that denervation polyuria is a transitory artefact, induced by unfavorable experimental conditions (anesthesia, etc.).

So far there is no proof that the renal nerves play any part in tubular reabsorption or excretion. On the other hand, the renal nerves can undoubtedly modify the tone of the afferent and efferent glomerular arterioles and thus act on glomerular filtration and renal blood flow.

The work of Trueta and his associates suggests that the innervation of the kidney may have great importance in certain normal and pathologic states, because when the renal nerves are stimulated a fundamental change in the distribution of renal blood occurs and cortical ischemia is produced. O'Connor and Verney[2] have shown that water diuresis is inhibited in the dog in certain emotional states and by electrical stimulation of the skin. Inhibition that takes place rapidly is mediated by the sympathetic; that which takes place more slowly is due to the release of antidiuretic hormone.

Diuresis

"Diuresis" means the excretion of urine, but it is generally used to signify an increase in the formation and excretion of urine. Diuresis, in the latter meaning of the term, can be provoked

[1] KLISIECKI, A., M. PICKFORD, P. ROTSCHILD, and E. B. VERNEY, *Proc. Roy. Soc., London, s.B.*, **112**, 521, 1933; THEOBALD, G. W., and E. B. VERNEY, *J. Physiol.*, **83**, 341, 1935.

[2] O'CONNOR, W. J., and E. B. VERNEY, *loc. cit.; ibid.*, **33**, 70, 1945.

in many ways, by administering water, saline solutions, and certain diuretic drugs.

Water diuresis. Ingestion of a more or less large amount of water (1 to 2 liters) provokes diuresis in man, beginning in less than ½ hr., reaching its peak in 1 or 2 hr., and returning to the initial rate in 3 to 5 hr. Usually more water is excreted than the amount taken in, and pale urine of low specific gravity is passed.

The cause of this diuresis is the sudden dilution of the blood, which is nevertheless not very marked and lasts only a short time, as the water rapidly passes into the interstitial fluid. The change that takes place, whatever it may be (decrease in the osmotic pressure, or in the total base, or in the concentration of certain ions, etc.), does not act directly on the kidneys, but through the hypophysis. Verney and his collaborators[1] have shown that the secretion of the antidiuretic hormone of the hypophysis is controlled by the osmotic pressure of the plasma and extracellular fluids. Changes in osmotic pressure stimulate osmoreceptors situated in the territory of the internal carotid artery, probably in the central nervous system (supraoptic nuclei, etc.). During water diuresis the antidiuretic hormone circulating in the blood diminishes, and the time necessary for this decrease explains the delay between maximal diuresis and maximal hydration of the blood. Intravenous injection of hypotonic salt solution also provokes diuresis. As in the case of diuresis following ingestion of water, the maximal rate of urine excretion is delayed with respect to the maximal dilution of the plasma. Verney refers to water diuresis as a "physiological diabetes insipidus."

Water diuresis can be inhibited in several circumstances which provoke the secretion of the antidiuretic hormone. Thus exercise and certain conditioned reflexes, when they occur in the course of water diuresis, diminish the excretion of urine within a few minutes, even in denervated kidneys. In both these cases the posterior lobe of the hypophysis is stimulated through the central nervous system and secretes a larger amount of antidiuretic hormone.

On the other hand, morphine and other nonvolatile anesthetics provoke oliguria and prevent water diuresis. Possibly this type of anesthesia suppresses a normal inhibitory mechanism that acts on the hypophysis and there is an excess secretion of the antidiuretic hormone.

[1] VERNEY, E. B., *Lancet*, **2,** 739 and 781, 1946.

Saline diuresis. Intravenous injection, or ingestion of concentrated saline solution, produces transitory and more or less marked diuresis. Two mechanisms are active in this type of diuresis, according to whether the substance is readily reabsorbed by the tubes (chloride, glucose) or is either not reabsorbed or reabsorbed only in a slight degree (urea, sulfate, nitrate, phosphate, mannitol). In the first case, *e.g.*, with a concentrated NaCl solution, there is hydremia due to the passage of fluid from the tissues into the blood. This plethora increases the blood pressure and renal blood flow, with a consequent increase in glomerular filtration; this alone can increase diuresis even if no change in tubular reabsorption takes place. In the second case (urea, sodium sulfate, sucrose, mannitol, etc.) the substances, on not being reabsorbed by the tubes, increase the osmotic pressure and retain water in the lumen, thus acting as diuretics. Acidotic salts (NH_4Cl, $CaCl_2$) perhaps act in the same way, but it is impossible to deny that their diuretic activity may be due to changes in the blood plasma. Polyuria in pancreatic diabetes and in phlorhizin diabetes is due, at least in part, to the high concentration of glucose in the tubes, which on not being totally reabsorbed acts by osmosis and prevents the reabsorption of water.

Diuretic drugs. Certain xanthine compounds (caffeine, theobromine, theophylline) have a moderate diuretic effect in man. The mechanism by which they act is not well known. They increase the circulating blood volume and also have a direct effect on the kidney (increased glomerular filtration, and perhaps diminished tubular reabsorption).

Mercurial diuretics (salyrgan, novasurol)[1] act directly on the kidney. Their main effect consists in diminishing tubular reabsorption of sodium and chloride, and the urinary excretion of these substances increases considerably. The excretion of water also increases, but not necessarily because of a direct effect of the diuretic on the mechanism of water reabsorption. Probably the increase in water excretion is due to the loss of sodium from the plasma, or to the osmotic effect of the electrolytes which are not reabsorbed in the proximal tube. Simultaneous administration of NH_4Cl reinforces the effect of mercurial diuretics.

[1] GOVAERTS, P., *Arch. internat. de pharmacodyn. et de thérap.*, **36,** 99, 1929.

Excretion of Electrolytes

The clearances of many substances have been measured. Some of these occur normally in the organism (creatinine, urea, phosphate, sulfate, uric acid, amino acids, chloride, sodium, etc.); others are foreign substances (inulin, diodrast, phenolsulfonphthalein, xylose, saccharose, sulfonamides, etc.). It would not be profitable to examine here all the results obtained. Nevertheless it is interesting to note that there is a maximum reabsorption rate of phosphate, amino acids, creatinine, and ascorbic acid, just as in the case of glucose. This maximum is not constant but changes from time to time, perhaps in response to hormonal influences. In normal conditions the kidney discriminates between Na^+ and K^+, between Cl^- and HCO_3^-, and between other anions and cations. Thus a constant ionic equilibrium is maintained in the plasma. Within certain limits one special ion can be retained if it is scarce, or can be excreted if its concentration in the plasma is excessive. However, the amount of water and the total electrolyte concentration in plasma have a greater influence on renal function than the relative amounts of the different ions and the hydrogen ion concentration.

The effect of the corticoadrenal hormones on renal excretion of electrolytes is well known. Tubular reabsorption of Na is in great part dependent on the effect of these hormones, which cause it to be reabsorbed. The effect of parathyroid hormone on the renal excretion of calcium and phosphate has been explained in the chapter on mineral metabolism.

In dogs submitted to certain experimental conditions (excess K, urea diuresis), the ratio between potassium and creatinine clearances is more than 1; therefore, the distal tubes can excrete potassium.[1] Barclay, Cooke, and Kenney[2] believe that urea, phosphate, potassium, and other substances are eliminated by a triple process of glomerular filtration, tubular reabsorption, and tubular secretion. Tubular secretion would then occur only when the concentration in plasma of these substances was exceedingly high.

[1] Berliner, R. W., and T. J. Kennedy, *Proc. Soc. Exper. Biol. & Med.*, **67**, 542, 1948.

[2] Barclay, J. A., W. T. Cooke, and R. A. Kenney, *Acta med. Scandinav.*, **128**, 500, 1947.

Functions of the Kidney in Acid-Base Equilibrium

A constant hydrogen ion concentration is an outstanding feature of blood plasma. The pH is kept around 7.4 by the buffer action of the salts of weak acids (H_2CO_3, HPr, H_3PO_4, etc.). Carbonic acid is the principal acid produced in the process of oxidation of foodstuffs. A normal subject has an output of over 20 gram molecules in 24 hr., the equivalent of 2 liters of concentrated HCl, and about twenty times the total available base in the body. The anhydride of this acid (CO_2) is volatile and is therefore excreted almost exclusively by the lung, but the concentration of $BHCO_3$ in the plasma—quite as important as H_2CO_3 in determining the pH of the plasma—is regulated not by the lung but by the kidney (see page 287).

Other acids besides H_2CO_3 are formed in both the normal and the diseased organism. These acids are not volatile, and therefore cannot be excreted by the lungs. The metabolism of 100 gm. of protein ingested daily produces 60 mEq. of sulfate by oxidation of the protein S, and about 50 mEq. of phosphate by oxidation of the protein P; the metabolism of 100 gm. of fat, with 10 per cent lecithin, produces another 50 mEq. of phosphate. In cases of severe ketosis, up to 500 mEq. of β-hydroxybutyric acid can be added. The production of acid exceeds that of base by 30 to 80 mEq. daily.

This excess acid is neutralized by the base in the body fluids, but if the acid were excreted combined with this base, the alkaline reserve would be rapidly exhausted. The kidney has two different mechanisms by which it can excrete acid and retain the base, especially plasma bicarbonate. One of these mechanisms consists in the excretion of free acid; the other in the substitution of the fixed base by NH_3, which is synthesized in the kidney from the N of amino acids (page 742). By means of the first mechanism 10 to 30 mEq. of acid is eliminated daily, and 20 to 50 mEq. by means of the second.

The kidney contributes to the regulation of the hydrogen ion concentration in the first place by excreting an acid or an alkaline urine. Usually the pH of human urine is around 6, but it can fall to 4.8 in cases of acidosis and rise to 8.2 in cases of alkalosis. The power to excrete acid urine (retaining fixed base) is dependent

on the available buffer. Normally this buffer consists mostly of phosphate. At pH 7.4 in blood plasma there is Na_2HPO_4 and NaH_2PO_4 in a ratio of 4:1. In urine at pH 4.8, 99 per cent of the phosphate is found as NaH_2PO_4; therefore there is some sparing of base. At this same pH all the sulfate is in the form of B_2SO_4 and 90 per cent of the lactate as B-lactate. The excretion of free acid therefore has little physiologic significance, but when β-hydroxybutyric acid is excreted in urine at pH 4.8, 55 per cent is free; in this case there is an effective economy of fixed base.

Three explanations have been proposed to account for the acidification of urine in the renal tubes: (*a*) reabsorption of dibasic phosphate by the tubes; (*b*) glomerular filtration of H_2CO_3 and subsequent reabsorption of $NaHCO_3$; (*c*) the exchange of H^+ for Na^+ as the glomerular filtrate passes down the tube, so that Na_2HPO_4 is converted into NaH_2PO_4. The source of H^+ would be the H_2CO_3 produced by the tubular cells from the CO_2 in the blood or in the course of cellular oxidation.[1]

The administration of the ammonium salts of mineral acids (*e.g.*, NH_4Cl) is used to provoke experimental or therapeutic acidosis. The ammonia is almost completely converted into urea, and the free Cl combines with base displaced from HCO_3, the resulting CO_2 being excreted by the lung. Mineral salts of Ca, Mg, and Sr also provoke acidosis because the basic ion is excreted in the feces as phosphate or carbonate or as soaps formed with fatty acids, and the acid ion (Cl^-, $SO_4^{=}$) is set free, so it displaces base from HCO_3^-. In both cases there is a shift toward the acid in the urinary pH. This same process occurs when there is a high-protein diet, which produces a large quantity of acid.

When the plasma contains an excess of fixed base, either because of ingestion of alkaline salts or because of a vegetarian diet, the kidney contributes to the regulation of the acid-base equilibrium of the *milieu intérieur* by forming an alkaline urine; bicarbonate replaces Cl^-, and the dibasic phosphate increases. In both ways Cl^- is retained and its excretion combined with the excess base is prevented.

The "alkaline tide." Leathes observed that urinary acidity diminishes during the hours following awakening and after a meal. The morning "alkaline tide" seems to be related to changes in respiratory rhythm caused by the waking state. Hyperpnea increases urinary alkalinity because of the excess elimination of CO_2 by the lung and the subsequent excretion of $BHCO_3$ by the kidney. The postprandial alkaline tide is attributed to the accumulation in the blood of base set free by the secretion of HCl in the gastric juice. The increase in base in the blood, counteracted in part by the excretion of the alkaline pancreatic and intestinal juices, causes a reduction in the excretion of acid and NH_3 by the kidney. When HCl is reabsorbed the urine again becomes acid. In cases of achlorhydria the postprandial alkaline tide cannot be demonstrated.

[1] PITTS, R. F., and R. S. ALEXANDER, *Am. J. Physiol.*, **144**, 239, 1945.

FORMATION OF NEW SUBSTANCES BY THE KIDNEY

Bunge and Schmiedeberg[1] discovered that the injection of benzoic acid (sodium benzoate) and glycine is followed by the accumulation of hippuric acid in the blood, but no hippuric acid is formed if the renal blood vessels are first tied off. The perfusion of a kidney with blood containing benzoate and glycine showed that this organ synthesizes hippuric acid. This observation, which has been repeatedly confirmed, proves that the kidney has a power of synthesis of metabolic importance. In this particular case it forms part of an antitoxic defensive mechanism. In carnivora, especially in the dog, the kidney alone performs this operation, but in herbivora and in omnivora, including man, the liver plays a more important part than the kidney, to such an extent that the rate of excretion of hippuric acid after the administration of benzoic acid has been used as a test of hepatic function.

The kidney can also synthesize ammonia, which is excreted as ammonium salts and plays an important part in the maintenance of the acid-base equilibrium. By combining with acid ions, NH_3 spares fixed bases (Na, K). When the ingestion of acids, or of foodstuffs that give acids as end products, increases, the urinary excretion of ammonia also increases. On the contrary, when base or base-forming foodstuffs are ingested, the production of ammonia diminishes. Ammonia is produced and excreted by the distal tube. Ammonia is excreted by sim-

[1] BUNGE, G., and O. SCHMIEDEBERG, *Arch. f. exper. Path. u. Pharmakol.*, **6**, 233, 1876.

ple diffusion, according to some workers; others maintain that there is active secretion of ammonia.

Urea is not the source of urinary ammonia, as was at one time believed. The renal cortex has a very strong deaminating activity. The principal precursors of urinary NH_3 seem to be glutamic acid and its amide, glutamine. The conversion of glutamine into NH_3 increases with the degree of acidosis. Glycerine, D-L-alanine, L-leucine, and D-L-aspartic acid can also increase renal production of ammonia. The capacity of an amino acid to increase the formation of ammonia by the kidney is related to the facility with which it is deaminated by renal enzymes *in vitro* and to the readiness with which it is reabsorbed by the renal tubes.[1]

Synthesis of hippuric acid and the formation of ammonia from amino acids are not the only processes of this type carried out by the kidneys. Recent work has shown that other substances are also produced. The secretion of renin will be discussed below.

The Work of the Kidney

In order to concentrate the urine and to excrete substances, the renal tubes perform work and consume energy. Renal work in man when excreting highly concentrated urine is equivalent to 0.6 cal./min. The oxygen consumption of the human kidneys is about 10 cc./min. (equivalent to 50 cal./min.) and is subject to only small variations in spite of considerable changes in the osmotic work of the kidney. The efficiency of the kidney even in extreme conditions, compared with that of other organs, is very low; it is not much more than 1 per cent.

The kidney not only excretes urine but has other important functions. Probably the greater part of the energy consumed is employed in cell processes not directly concerned with external osmotic work.

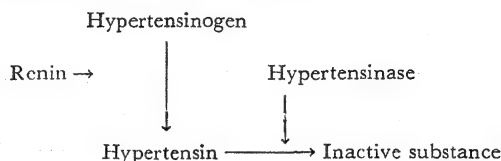
ARTERIAL HYPERTENSION OF RENAL ORIGIN

Renin has been mentioned as a substance formed by the kidney and secreted into the blood. Renin is a protein found in renal extracts,[2] which produces an increase in blood pressure when it is injected into the blood stream. The increase in pressure is not due to the direct action of renin on the diameter of the blood vessels. Renin is not per se a vasoconstrictor, it is a proteolytic enzyme which acts on a serum globulin (hypertensinogen) and splits off a simpler substance, possibly a polypeptide, hypertensin, which constricts the blood vessels and stimulates the heart. Hypertensin is destroyed in the body by another enzyme or group of enzymes, called hypertensinase.

Hypertensinogen

Renin → | Hypertensinase

↓ ↓

Hypertensin ——→ Inactive substance

Renin in certain experimental conditions is secreted into the blood; therefore it can be considered as an internal secretion. The kidney secretes renin in two different conditions:

1. In arterial hypotension provoked by (*a*) the injection of substances that lower the blood pressure; (*b*) hemorrhage; (*c*) shock;[1] etc.
2. When the renal blood flow decreases (partial occlusion of the renal artery, perinephritis, ureteral occlusion, etc.) or there is total suppression of the renal circulation (temporary occlusion of the renal artery).

In the first case, when the general blood pressure falls, renin secretion is one of the mechanisms of homeostasis. In the same way as arterial hypotension provokes reflex vasoconstriction and adrenaline secretion, which help to restore the normal blood pressure level, so it also provokes renin secretion, which by the formation of hypertensin contributes to raise the lowered blood pressure. The mechanism by which it provokes the secretion, however, is unknown.

In the second case, *i.e.*, when the renal circulation is diminished, the renin secreted causes a pathologic syndrome: arterial hypertension. The hemodynamic conditions of the kidney are probably the same in both cases; the renal blood flow diminishes, the pressure in the renal arteries falls, and the pulse pressure is less. In the first case, however, renin secretion has a physiologic function to perform; it helps to restore the normal blood pressure level. In the second case the general blood pressure is not below normal;

[1] Lotspeich, W. D., and R. F. Pitts, *J. Biol. Chem.*, **168**, 611, 1947.

[2] Tigerstedt, R., and P. G. Bergman, *Skandinav. Arch. f. Physiol.*, **8**, 223, 1898.

[1] Huidobro, F., and E. Braun-Menéndez, *Am. J. Physiol.*, **137**, 47, 1942.

therefore the secreted renin causes the abnormal syndrome of arterial hypertension. Probably in arterial hypertension of renal origin there are other factors beside those mentioned here, but undoubtedly in the first stages of experimental hypertension of renal origin,[1] in acute experimental hypertension, and in certain cases of hypertension in man (acute nephritis, eclampsia, etc.)[2] renin plays an important part.

Since Bright's classic observations, clinicians had associated arterial hypertension with renal damage, and many efforts were made to reproduce the disease in laboratory animals. Many unsuccessful attempts were made (surgical reduction of the renal mass; incomplete ligature of the renal vein; parenchymatous lesions of the kidney by x-rays, toxic substances, etc.) before Goldblatt and his collaborators[3] discovered that an incomplete obstruction of the renal artery provoked permanent hypertension in the dog.

Hypertension of renal origin, produced by these and similar methods, has many points of resemblance to human "essential hypertension." This simple method of obtaining experimental hypertension encouraged several groups of workers to search for the mechanism of the disturbance. It was soon demonstrated that this type of hypertension was not due to a nervous mechanism, as it was not prevented by renal denervation. The different endocrine glands were also eliminated as causes of the condition. A humoral mechanism therefore seemed to be its probable cause.

Houssay and Fasciolo[4] proved that a humoral mechanism existed. They transplanted the ischemic kidneys (with partially occluded renal arteries) of dogs with experimental hypertension into the neck of nephrectomized dogs, which thereupon showed an increase in blood pressure (Figs. 304 and 305). It is now known that this rise in blood pressure is due to the renin secreted by the ischemic kidney.

Houssay and Taquini[5] observed that the citrated plasma obtained from the venous blood of the ischemic kidney has a vasoconstrictor effect on the blood vessels of the toad (Läwen-Trendelenburg method).[1] This effect is now known to be produced by hypertensin formed in the blood by the action of renin on the hypertensinogen of blood plasma. Later it was seen that acute partial ischemia has an effect similar to that of chronic ischemia; it increases the blood pressure, and the renal venous blood plasma acquires a vasoconstrictor action. Moreover intravenous injection of the venous blood of kidneys in acute partial ischemia into normal dogs causes in the latter a rise in blood pressure.[2] Shortly afterward extracts of this blood were prepared and a pressor substance, which was called "hypertensin," was discovered[3] (Fig. 306). This substance could be formed *in vitro* by incubating blood plasma or serum with renin at 37°C. (Fig. 307). Hypertensin is formed in the blood by the action of renin, secreted by the kidney, on the hypertensinogen of plasma.

Page and his collaborators, following another path, arrived almost simultaneously at the same conclusions. In 1938[4] they observed that purified renin did not produce peripheral vasoconstriction (in the rabbit's ear or the dog's tail); the vasoconstrictor effect was reestablished by adding a plasma protein which they called "renin activator." It was then discovered that the interaction of renin and renin activator gave rise to a pressor and vasoconstrictor substance which was called "angiotonin."[5] There is therefore only a difference in the names given to the same principles by the two groups of workers.

Renin. This substance is a protein found in renal extracts, which acts on hypertensinogen and forms hypertensin. This is an enzymatic process, with an optimum pH of 7.5 to 8.5 and an optimum temperature of 37 to 39°C. The velocity of the reaction is dependent on the renin concentration, a fact that has been used for the

[1] Dell'Oro, R., and E. Braun-Menéndez, *Rev. Soc. argent. de biol.*, **18**, 65, 1942.

[2] Dexter, L., and F. W. Haynes, *Proc. Soc. Exper. Biol. & Med.*, **55**, 288, 1944.

[3] Goldblatt, H., J. Lynch, R. F. Hanzal, and W. W. Summerville, *J. Exper. Med.*, **59**, 347, 1934.

[4] Houssay, B. A., and J. C. Fasciolo, *Bol. Acad. nac. de med. de Buenos Aires*, p. 342, 1937; *Rev. Soc. argent. de biol.*, **13**, 284, 1937; *Compt. rend. Soc. de biol.*, **127**, 147, 1938.

[5] Houssay, B. A., and A. C. Taquini, *Rev. Soc. argent. de biol.*, **14**, 5, 1938; *Compt. rend. Soc. de biol.*, **128**, 1125, 1938.

[1] Fasciolo, J. C., B. A. Houssay, and A. C. Taquini, *J. Physiol.*, **94**, 28, 1938.

[2] Braun-Menéndez, E., and J. C. Fasciolo, *Rev. Soc. argent. de biol.*, **15**, 161 and 401, 1939.

[3] Braun-Menéndez, E., J. C. Fasciolo, L. F. Leloir, and J. M. Muñoz, *Rev. Soc. argent. de biol.*, **15**, 420, 1939: *J. Physiol.*, **98**, 283, 1940.

[4] Kohlstaedt, K. G., O. M. Helmer, and I. H. Page, *Proc. Soc. Exper. Biol. & Med.*, **39**, 214, 1938.

[5] Page, I. H., and O. M. Helmer, *J. Exper. Med.*, **71**, 29, 1940.

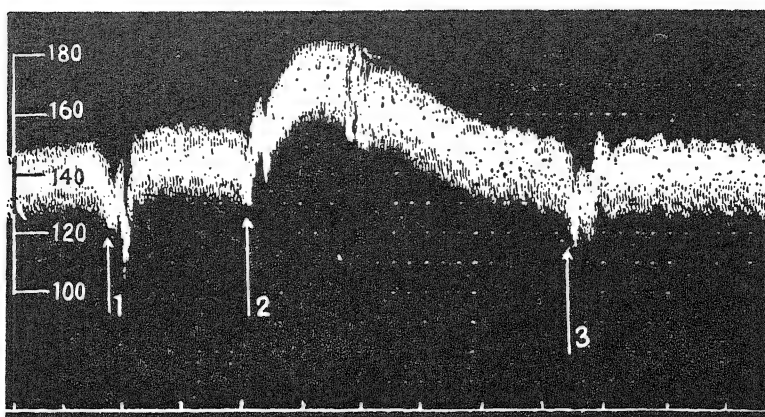



FIG. 306. Hypertensin in venous blood from an ischemic kidney. Femoral arterial blood pressure (in mm. Hg) of a vagotomized, chloralosed dog. Injection of acetone extract of 50 cc. serum of (1) blood from heart-lung preparation after 1 hr. circulation; (2) venous blood from an ischemic kidney in heart-lung-kidney preparation; (3) blood from heart-lung preparation after 3 hr. circulation. Time in minutes. (*Braun-Menéndez, E., J. C. Fasciolo, L. F. Leloir, and J. M. Muñoz, J. Physiol., vol. 98, p. 283, 1940.*)

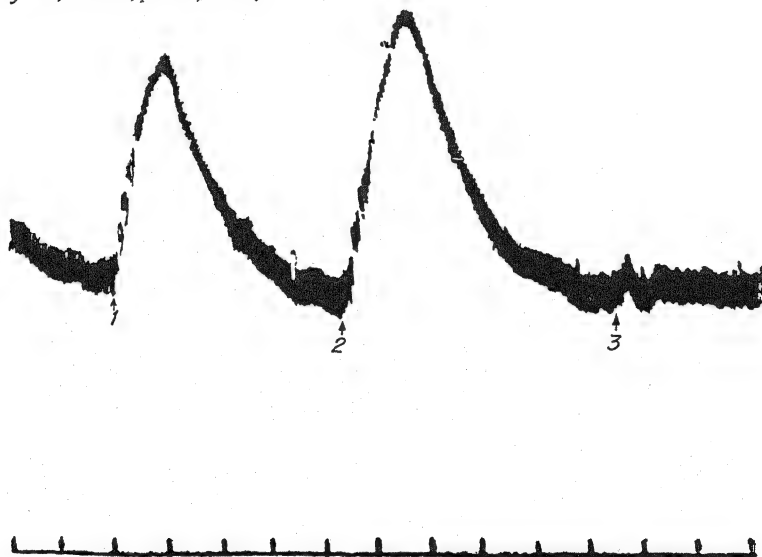



FIG. 307. Formation of hypertensin in vitro. Carotid pressure in the dog. Time in minutes. 1, intravenous injection of 1 cc. (1 unit) standard hypertensin solution; 2, injection of alcoholic extract of a mixture of 6 cc. bovine blood plasma and 1 cc. renin incubated at 37°C. during 5 min.; 3, the same without incubation.

quantitative estimation of renin.[1] Renin can be considered as a proteolytic enzyme,[2] but it is difficult to identify it as one of the known cellular proteases.[3]

The renins obtained from different species behave differently. The renins of man, monkey, and baboon are active on the hypertensinogens of all mammals and give rise to hypertensin, but the renins of the latter do not act on the hypertensinogen of the former.[1] The renin of birds is active only on the hypertensinogen of birds. No renin has been found in amphibians, reptiles, and fishes.[2]

[1] LELOIR, L. F., J. M. MUÑOZ, E. BRAUN-MENÉNDEZ, and J. C. FASCIOLO, *Rev. Soc. argent. de biol.*, **16**, 635, 1940.
[2] BRAUN-MENÉNDEZ, E., J. C. FASCIOLO, L. F. LELOIR, and J. M. MUÑOZ, *J. Physiol.*, **98**, 282, 1940.
[3] PLENTL, A. A., and I. H. PAGE, *J. Biol. Chem.*, **155**, **363**, 1944.

[1] FASCIOLO, J. C., L. F. LELOIR, J. M. MUÑOZ, and E. BRAUN-MENÉNDEZ, *Science*, **92**, 554, 1940.
[2] BEAN, J. W., *Am. J. Physiol.*, **136**, 731, 1942.

Intravenous injection of renin provokes an increase in blood pressure which lasts 30 to 40 min. This increase is due to the hypertensin formed by the action of renin on the hypertensinogen in the blood. Repeated injections of renin produce a progressively smaller response (tachyphylaxis), in great part because of the exhaustion of the hypertensinogen (Fig. 308).

Pepsin acting on hypertensinogen produces a substance called "pepsitensin,"[1] which has pharmacologic and chemical properties similar to those of hypertensin.[2] Nevertheless it is possible to differentiate these substances by means of the enzymes that destroy them.[3]

140
100
60
20
1
2
3
4

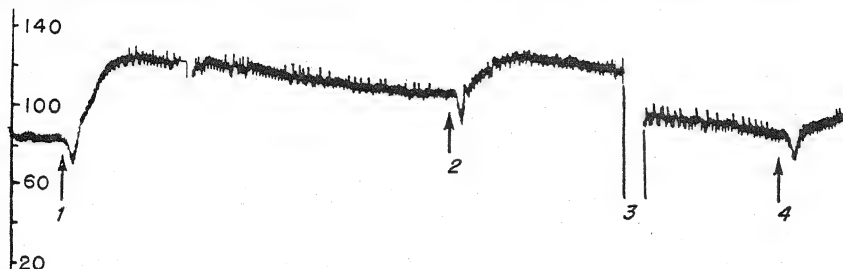



Fig. 308. Pressure action of repeated injections of renin. Carotid pressure in the dog. On the left, pressure scale in mm. Hg. Renin (1 cc.) was injected at (1), (2), and (4). Interval between (1) and (2) was 13 min.; at (3) the record was interrupted for 15 min. Note the decrease in the effect of the second and third injections. (*Braun-Menéndez, E., J. C. Fasciolo, L. F. Leloir, J. M. Muñoz, and A. C. Taquini, "Renal Hypertension," translated by Lewis Dexter, Charles C. Thomas, Springfield, Ill., 1946.*)

Hypertensinogen. This substance is one of the plasma proteins, and it forms part of the α globulins.[1] It is formed in the liver,[2] and the adrenal cortex seems to condition the velocity of its formation. In some cases of hypertension with renal insufficiency in man, the concentration of hypertensinogen in the blood has been found to be increased, and it has been found below normal in cases of hepatic insufficiency.[3] Page and his collaborators have proposed the name of "renin substrate α-2-globulin" for this substance.

Hypertensin (angiotonin). This substance is probably a polypeptide of relatively low molecular weight (approximately 2,700), which is produced by the action of renin on hypertensinogen. It is dialyzable, thermostable, and acid-resistant, and it is destroyed by alkalies and proteolytic enzymes (pepsin, trypsin, papain). It is soluble in water but is insoluble in the organic solvents.[4]

[1] Plentl, A. A., I. H. Page, and W. W. Davis, *J. Biol. Chem.*, **147**, 143, 1943; Cohn *et al.*, *J. Clin. Investigation*, **23**, 417, 1944.

[2] Page, I. H., B. McSwain, G. M. Knapp, and W. D. Andrus, *Am. J. Physiol.*, **135**, 213, 1941; Leloir, L. F., J. M. Muñoz, E. Braun-Menéndez, J. C. Fasciolo, and A. C. Taquini, *Rev. argent. de cardiol.*, **9**, 269, 1942.

[3] Haynes, F. W., and L. Dexter, *Federation Proc.*, **2**, 20, 1943.

[4] Braun-Menéndez, E., J. C. Fasciolo, L. F. Leloir, and J. M. Muñoz, *J. Physiol.*, **98**, 282, 1940; *Rev. Soc. argent. de biol.*, **16**, 398, 1940.

Hypertensin has a strong vasoconstrictor action. In intravenous injection it produces a brief increase in blood pressure in amphibians, reptiles, birds, and mammals (Figs. 306 and 307). Intravenous infusion produces a rise in blood pressure which lasts as long as the infusion is continued (Fig. 309). The pressor effect of hypertensin is not inverted by sympatholytic drugs, nor is it reinforced by cocaine. It is potentiated by veritol, tyramine, and ephetonine. It has a stimulating effect on the heart and produces contraction in most plain muscles. It has been classified among the musculotropic drugs, as it seems to act directly on the muscle fiber.[4]

Hypertensinase. Hypertensin is destroyed by serum, plasma, red blood cells, and tissue extracts (kidney, liver, spleen, intestinal mucosa, etc.). This effect is due to a thermolabile nondialyzable substance which has the properties of an enzyme and which has been called hypertensinase.[5] Several of the known enzymes destroy hypertensin.[6] The activity of tissue extracts is

[1] Croxatto, H., and R. Croxatto, *Science*, **95**, 101, 1942.

[2] Alonso, O., R. Croxatto, and H. Croxatto, *Proc. Soc. Exper. Biol. & Med.*, **52**, 61, 1943.

[3] Braun-Menéndez, E., J. C. Fasciolo, L. F. Leloir, J. M. Muñoz, and A. A. Taquini, *Rev. Soc. argent. de biol.*, **19**, 304, 1943; Plentl and Page, *op. cit.*, p. 379.

[4] Ludueña, F. P., *Rev. Soc. argent. de biol.*, **16**, 138, 1940.

[5] Leloir, Muñoz, Braun-Menéndez, and Fasciolo, *op. cit.*, 75; *Compt. rend. Soc. de biol.*, **134**, 487, 1940.

[6] Croxatto, H., *Sev. de med. y aliment.*, **5**, 259, 1943.

probably due to proteolytic enzymes that destroy hypertensin by hydrolysis.

The experimental study of hypertension of renal origin has influenced our ideas on the pathogenesis of human arterial hypertension, a disease that because of its frequency and severity is one of the major problems of medicine. These experimental studies have contributed directly and indirectly to a better understanding of this disease; nevertheless there remain many obscure points which require further research. The renal origin of those cases of hypertension which are cured by extirpation of the diseased kidney cannot be denied. In cases of acute hypertension with renal lesions (acute glomerulonephritis and toxemia of pregnancy) the presence of renin in the blood has been demonstrated. In cases of chronic hypertension there are also renal lesions, but it has not been possible to demonstrate an increase in the renin concentration in blood, either in experimental animals or in man. This is a weak point in the hypothesis that considers the renin-hypertensin mechanism as the cause of hypertension. Several explanations of this fact can be suggested: (*a*) once the hypertension is established, the amount of renin needed to maintain the high blood pressure may be so small that it cannot be detected by the methods so far available; (*b*) once hypertension is established, central, peripheral, and reflex nervous mechanisms may contribute to maintain the high blood pressure level.

Very small amounts of renin can be detected in the blood by means of highly sensitive methods recently developed.[1] Small amounts of renin have thus been detected in the circulating blood of normal and hypertensive dogs and rats; also in the blood of normal human subjects and of patients suffering from chronic hypertension. The same renin concentration has been found, however, in the blood of normal and hypertensive animals and men.[1] These facts show that vasoconstriction in human or experimental hypertension is not due to an increase in blood-renin concentration, but they do not completely discard renin as a factor in the pathogenesis of hypertension, and other experimental evidence shows that it can play a part in this process. Prolonged hypertension can be produced in rabbits by continuous injection of very small amounts of renin.[2] Repeated injections of heterologous renin into dogs have provoked the appearance of antirenin with a high titer in the blood plasma;[3] in hypertensive dogs the blood pressure falls when antirenin appears. In dogs with antirenin in their blood, occlusion of the renal arteries has not been followed by hypertension. Renin also has a diuretic effect[4] and causes proteinuria and loss of sodium in the urine. It also provokes hypertrophy of the glomerular zone of the adrenals.

The second hypothesis referred to above is supported by certain experimental evidence, which shows that a high blood pressure may be maintained by nervous and humoral mechanisms

FIG. 309. Continuous intravenous injection of hypertensin at the rate of 1 unit per minute. 1, injection started; 2, injection ended. Chloralosed, vagotomized dog with artificial respiration. Blood pressure in mm. Hg. Time in minutes. (*Braun-Menéndez, E., J. C. Fasciolo, L. F. Leloir, and J. M. Muñoz, J. Physiol., vol. 98, p. 283, 1940.*)

[1] FASCIOLO, J. C., and A. C. TAQUINI, *Rev. Soc. argent. de biol.*, **23**, 138, 1947.

[1] TAQUINI, A. C., and FASCIOLO, J. C., *Rev. argent. cardiol.*, **1**, 14, 1947; BRAUN-MENÉNDEZ, E, M. R. COVIÁN, and C. E. RAPELA, *Rev. Soc. argent. de biol.*, **23**, 131, 1947.

[2] BLACKET, R. B., *et al.*, *Clin. Sc.*, **9**, 223, 1950.

[3] BURNS, R. O., JR., and G. E. WAKERLIN, *Proc. Central Soc. Clin. Research*, 24, 20, 1951.

[4] MASSON, G. M. C., A. C. CORCORAN, and I. H. PAGE, *J. Lab. & Clin. Med.*, **38**, 213, 1951; CROXATTO, H., L. BARNAFI, and J. PASSI, *Acta physiol. latinoam.*, **2**, 159, 1952.

not dependent on the action of renin. Thus a substance called vasoexcitor material (VEM) arising in the kidney has recently been studied, but it is not yet possible to say what part it plays in the pathogenesis of arterial hypertension. Chambers, Zweifach, and their associates[1] have followed the reactions of the blood vessels in the exteriorized mesoappendix of rats submitted to shock. They were able to distinguish two phases: (*a*) a first period during which there is compensation, the meta-arterioles and the precapillary sphincters are hyperactive, and there is an increase in the response to adrenaline; (*b*) a final phase of progressive decompensation in which the activity of the terminal arterioles and precapillaries is depressed, and their response to adrenaline is decreased. Blood taken from animals in these conditions was injected into normal rats with the mesoappendix exposed, and the same vascular responses were provoked as were observed in the donor. They conclude that there are two substances in the blood of animals in shock: (*a*) vasoexcitor material (VEM), which appears during the first stages of shock; (*b*) vasodepressor material (VDM), which appears after prolonged tissue anoxia and rises to a high concentration at the beginning of the irreversible phase of shock. VEM is produced in the kidney and VDM in the liver and muscles submitted to severe anoxia.[2] These substances are apparently part of a homeostatic mechanism that contributes to the regulation of blood pressure and peripheral blood flow. VEM has certain similarities with hypertensin, although some of its vascular effects differ from those of the latter. It is destroyed by kidney tissue *in vitro* in aerobic conditions, and it appears in the blood of the renal vein after a few minutes of acute partial ischemia. The blood of dogs with constriction of the renal artery has a strong vasoexcitatory effect, but when the blood pressure is stabilized at a high level the blood has little or no excitatory effect.[3] This "neutral effect" of the blood may be due not to a fall in VEM but to a rise in VDM, which has been

[1] ZWEIFACH, B. W., R. E. LEE, C. HYMANS, and R. CHAMBERS, *Ann. Surg.*, **120**, 232, 1944; ZWEIFACH, B. W., R. G. ABELL, R. CHAMBERS, and G. H. CLOWES, *Surg., Gynec. and Obst.*, **80**, 593, 1945.

[2] SHORR, E., B. W. ZWEIFACH, and R. F. FURCHGOTT, *Science*, **102**, 489, 1945.

[3] ZWEIFACH, B. W., E. SHORR, and S. BAEZ, *Federation Proc.*, **6**, 232, 1945.

identified with ferritin.[1] Blood from hypertensive dogs or patients with chronic hypertension incubated with slices of kidney in aerobic conditions, so as to destroy VEM, shows VDM activity. On the other hand, antibodies for VDM have been prepared which counteract its effects and reveal the activity of VEM when added to blood from hypertensive animals or patients. Blood taken from animals with experimental hypertension and patients with chronic hypertension treated in this way consistently revealed appreciable amounts of VEM and VDM. Blood of normal animals and men was found to be inactive before and after incubation with kidney tissue. In chronic arterial hypertension in man and dogs, therefore, there is an increase in VEM masked by an equivalent increase in VDM.[2]

These observations confirm and extend the role of the kidney as an organ secreting substances with vascular activity and increase its importance in the pathogenesis of hypertension.

BIBLIOGRAPHY

BARCROFT, J., "Features in the Architecture of Physiological Function," Cambridge, New York, 1934.

BERNARD, C., "Leçons sur les propriétés des tissus vivants," J. C. Baillière et fils, Paris, 1866.

BERNARD, C., "Lecons sur les phénomènes de la vie communs aux animaux et aux végétaux," J. C. Baillière et fils, Paris, 1878.

CANNON, W. B., "The Wisdom of the Body," Norton New York, 1932.

DAREMBERG, C., "Oeuvres anatomiques, physiologiques et médicales de Galien," vol. 1, J. C. Baillière et fils, Paris, 1854, p. 353.

GOLDBLATT, H., The Renal Origin of Hypertension, *Physiol. Rev.*, **27**, 120, 1947.

HALDANE, J. S., "Respiration," Yale University Press, New Haven, 1922.

HALDANE, J. S., "The Philosophical Bases of Biology," Hodder, London, 1931.

HENDERSON, L. J., "Blood. A Study in General Physiology," Yale University Press, New Haven, 1928.

HERVIN, R. C., Factors Affecting the Tests of Kidney Function, *Physiol. Rev.*, **21**, 529, 1941.

MARSHALL, E. K., The Secretion of Urine, *Physiol. Rev.*, **6**, 440, 1926.

MARSHALL, E. K., The Comparative Physiology of the Kidney in Relation to Theories of Renal Secretion, *Physiol. Rev.*, **14**, 133, 1934.

[1] MAZUR, A., and E. SHORR, *J. Biol. Chem.*, **176**, 771, 1948.

[2] SHORR, E., B. W. ZWEIFACH, and S. BAEZ, *Federation Proc.*, **7**, 115, 1948; SHORR, E., and B. W. ZWEIFACH, *Federation Proc.*, **8**, 1, 1949.

PETERS, J. P., and D. D. VAN SLYKE, "Quantitative Clinical Chemistry," 2d ed., Williams & Wilkins, Baltimore, 1946.

PITTS, R. F., Renal Excretion of Acid, *Federation Proc.*, **7**, 418, 1948

RICHARDS, A. N., "Methods and Results of Direct Investigation of the Function of the Kidney," Williams & Wilkins, Baltimore, 1929.

RICHARDS, A. N., *Harvey Lect.*, **30**, 93, 1934–1935.

RICHARDS, A. N., Croonian Lecture, *Proc. Roy. Soc., London, s.B.*, **126**, 398, 1938.

SHANNON, J. A., Renal Tubular Excretion, *Physiol. Rev.*, **19**, 63, 1939.

SMITH, H. W., "The Physiology of the Kidney," Oxford, New York, 1937.

SMITH, H. W., The Excretion of Water, *Bull. New York Acad. Med.*, **23**, 177, 1947.

SMITH, H. W., "The Kidney. Structure and Function in Health and Disease," Oxford, New York, 1951.

SMITH, H. W., Physiology of the Renal Circulation, *Harvey Lect.*, **35**, 166, 1939–1940.

VERNEY, E. B., The Antidiuretic Hormone and the Factors Which Determine Its Release, *Proc. Roy. Soc., London, s.B.*, **135**, 25, 1947.

WINTON, F. R., Physical Factors Involved in the Activities of the Mammalian Kidney, *Physiol. Rev.*, **17**, 408, 1937.

CHAPTER 64

The Physiology of Micturition

URINE FORMED IN the kidney passes out of the orifices of the cribriform area into the renal calyces and pelvis. It then descends along the ureter to the bladder, where it accumulates. Periodically the bladder is emptied by the act of micturition.

The renal pelvis and the ureter are made up of three layers: an external connective adventitia, a middle muscular coat, and an internal mucous membrane. The muscular coat consists of two layers of plain muscle—a deep one of longitudinal fibers, surrounded by another of circular fibers; in the lower part of the ureter there is an outside layer of longitudinal fibers. The ureter penetrates obliquely into the bladder wall, opening into its cavity (ureteral meatus) in such a way that a fold is formed which acts as a valve and prevents the reflux of urine from the bladder to the ureter. Rhythmic contractions of the pelvic and ureteral muscle are observed. They are propagated from the pelvis down the ureter in the form of a peristaltic wave at the rate of 2 to 3 cm. per sec.

These peristaltic waves are responsible for moving the urine along the ureter; the rhythmic ejaculation of a few drops of urine from the ureteral meatus can be directly observed by cystoscopy. These waves recur with a frequency of one to five per minute in man. Their amplitude and frequency increase when the ureter is distended by a greater volume of urine, by resistance to flow caused by a mechanical obstacle, or by a rise in temperature.

The ureter is innervated by sympathetic fibers from the renal, hypogastric, and spermatic plexuses. The lower third of the ureter also receives fibers from the pelvic parasympathetic. In the lower part of the ureter ganglionic neurons are found, but not in the upper part. The peristaltic waves of the ureter are not dependent on the integrity of its innervation; the ureter, like the heart, has an automatic rhythm. The ureter contracts rhythmically when isolated in the same way as it does *in situ*.

The bladder and the prostatic and membranous segments of the urethra are one functional unit, the object of which is (*a*) the accumulation in the bladder of urine coming from the kidneys; (*b*) its periodic evacuation.

The bladder has a coat of plain-muscle fibers, disposed in three layers. The contraction of these fibers (detrusor urinae) reduces the capacity of the bladder. The outside layer is made up of longitudinal fibers going from the fundus to the base of the bladder; the anterior fibers end by insertion on the pubis, and the posterior fibers on the prostate in the male and on the connective tissue of the urethrovaginal cleft in the female. The internal or plexiform layer is formed by longitudinal fibers, which anastomose with each other. The middle layer consists of circular fibers, forming an annular muscle around the opening of the urethra which extends into the depth of the prostate in man; this is the vesical or internal sphincter of the urethra. Urine is retained in the bladder by the involuntary tonic contraction of the smooth fibers of this sphincter. A coat of striated muscular fibers surrounds the prostatic and membranous segments of the urethra, forming the external sphincter. Voluntary contraction of this sphincter constricts the urethra and reinforces the retaining effect of the internal sphincter. By its contraction this muscle also empties the upper segment of the urethra at the end of micturition. The same effect is produced by the contraction of the urethral portion of the bulbocavernous muscle, an effect that has given it the name "accelerator urinae."

Innervation of the bladder. (Fig. 310.) The muscles of the bladder wall and the internal sphincter (plain muscle) have a double innervation: they receive sympathetic fibers from the hypogastric nerves (hypogastric plexus in man) and parasympathetic fibers from the pelvic nerves (nervi erigentes). The striated external sphincter is innervated by a perineal branch of the internal pudic nerve. The sympathetic fibers come from neurons in the lateral gray column of the upper lumbar segments of the spinal cord; their axons pass into the lumbar sympathetic chain and the mesenteric plexus. The majority of the fibers join the presacral nerve (upper hypogastric plexus), from which the hypogastric nerves branch off to end in the hypogastric ganglion situated on each side of the rectum. The neurons that give off the postganglionic fibers for the bladder are in these ganglia. Stimulation of the sympathetic provokes vasoconstriction, but there is no evidence that it has any other action on the bladder.

FIG. 310. Innervation of the bladder by the inferior mesenteric and hypogastric plexuses. (*White, J. C., and R. H. Smithwick, "The Autonomic Nervous System,"* 2d *ed., Macmillan, New York, 1945.*)

The parasympathetic fibers arise in the second to fourth sacral segments of the spinal cord; they join the pelvic nerves and end on neurons situated in the hypogastric ganglia or in the bladder wall. The axons of these neurons are the postganglionic fibers that end in the muscle fibers of the bladder. Stimulation of the parasympathetic contracts the bladder muscle (detrusor urinae) and relaxes the internal sphincter. Both these actions serve to empty the bladder. Acetylcholine has the same effect.

Afferent impulses from the bladder are conducted almost exclusively by fibers in the pelvic nerve. These fibers enter the spinal cord by the dorsal roots of the second and third sacral segments. The fibers that conduct painful impulses enter the spinal cord in the dorsal roots of the upper two or three lumbar nerves.

The pudic nerves innervate the external sphincter. They come from the three upper sacral nerves. They also carry the afferent fibers from the urethra.

The filling of the bladder. Urine is propelled into the bladder by the rhythmic peristaltic contractions of the ureters, gradually filling it. The bladder muscle, in the same way as the muscular coat of other hollow viscera, can adapt its capacity to the contents without any considerable change in the intravesical pressure. This postural adaptation reflex can be demonstrated by catheterizing the bladder and connecting the catheter to a manometer. If fluid is then injected into the bladder, 50 cc. at a time, the injection of the first 400 cc. produces only slight and transitory changes in pressure. This fact proves that the detrusor presents only a momentary tonic resistance to distention and then adapts its tonus to the content. The pressure nevertheless rises gradually as fluid is injected into the bladder (Fig. 311). When the contents of the bladder are more than 400 cc., the pressure curve rises more steeply, especially when the contents reach 600 to 800 cc. Rhythmic, painful contractions commence at this level. Micturition can be prevented by a voluntary effort, but when the pressure is above 100 cm. H_2O the bladder is evacuated however strong the will to retain the urine.

Retention of urine in the bladder. Urine is retained in the bladder by the closure of the ureteral meatuses, which prevents its returning to the kidney, and by the tonic contraction of the sphincters, which prevents its flowing into the urethra. The oblique opening of the ureter, the fold formed by the upper lip of the meatus, and the contraction of the bladder wall ensure the closure of the ureteral meatus. The sphinc-

ters resist a pressure of 70 to 100 cm. H_2O. This resistance is due to a reflex contraction of the internal (plain-muscle) sphincter, adapted to the volume of urine in the bladder (postural reflex) and to automatic or voluntary contraction of the external (striated-muscle) sphincter. When the pressure rises to 15 or 18 cm. H_2O, as normally happens when the contents are about 400 cc., afferent impulses are sent to the nerve centers and a desire to micturate is felt.

Micturition. Micturition is a reflex act, but normally it is commenced voluntarily. A purely involuntary micturition takes place only in children and in cases where the bladder has been deprived of its connections with the higher nerve centers. When the desire to micturate is felt, evacuation is prevented or performed voluntarily. In the first case voluntary contraction of the external sphincter reinforces the effect of the internal sphincter and there is reciprocal relaxation of the detrusor; the intravesical pressure diminishes and the desire to micturate ceases for a time. The desire to evacuate the bladder can be voluntarily overcome as long as the bladder contents are less than 700 cc.; there is then a painful sensation in the hypogastrium and urgent need to evacuate. When the desire to micturate is complied with, the voluntary contraction of the external sphincter is inhibited and a chain of reflexes commences, which results in the emptying of the bladder. Nevertheless micturition can be interrupted voluntarily by contraction of the external sphincter.

During micturition the detrusor contracts, the internal sphincter is relaxed, and urine is evacuated with considerable force; the pressure within the bladder can rise up to 130 cm. H_2O. Micturition is usually preceded by contraction of the abdominal muscles and a slight expiratory effort with the glottis closed. This effort assists micturition but is not indispensable to it. Abdominal contraction can be sustained throughout the act to accelerate evacuation. In the male the bulbocavernous muscle contracts rhythmically at the end of micturition, and thus the last drops of urine in the urethra are expelled (Janet's "piston stroke"). In the female micturition ends more abruptly.

Reflex mechanism of micturition. Barrington's[1] work in the cat has demonstrated the existence of a series of coordinated reflexes in the act of micturition:

1. Contraction of the bladder wall (detrusor) when the internal pressure reaches 10 cm. H_2O. The afferent and efferent paths of this reflex are in the pelvic nerve; the reflex center is situated in the pons and the medulla.

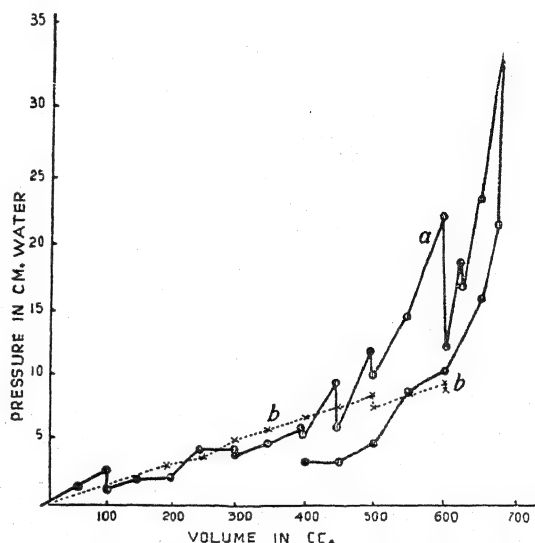



FIG. 311. Intravesical pressure: *a*, solid line, record of pressure following repeated injections of 50 cc. at short intervals; the vertical drop indicates postural adaptation of the bladder during these intervals; *b*, dotted line, record of pressure variations in a similar experiment made after the bladder was submitted to prolonged distention. (*Denny-Brown, D., and E. G. Robertson, Brain, vol. 56, pp. 149 and 397, 1933.*)

2. Contraction of the bladder wall, provoked by the flow of urine into the urethra. This reflex ensures the complete evacuation of the bladder once it has commenced. Its afferent fibers form part of the pudic nerve, the efferent path is in the pelvic nerve, and the center is situated in the pons and the medulla.
3. Weak, transitory contraction of the bladder caused by distention of the first part of the urethra. The afferent and efferent paths of this reflex are in the hypogastric nerve.
4. Relaxation of the external sphincter when urine flows through the urethra. The pudic nerve carries the afferent and efferent paths of this reflex.
5. Relaxation of the sphincter provoked by the contraction of the bladder wall. The afferent

[1] BARRINGTON, F. J. F., *Quart. J. Exper. Physiol.*, **8**, 33, 1914; **9**, 261, 1915; **15**, 81, 1925; *Brain*, **44**, 23, 1921; **45**, 126, 1922; **51**, 209, 1928; **54**, 177, 1931.

path is in the pelvic nerve and the efferent path in the pudic nerve.

6. Relaxation of the plain muscle of the upper third of the urethra when the bladder is distended. The afferent and efferent paths of this reflex are in the pelvic nerve.

The centers of the last four reflexes are situated in the sacral segments of the spinal cord.

In the course of micturition each one of these reflexes, with the exception of the third, provokes the following one, and thus the bladder is completely evacuated.

The nervous system and the functions of the bladder. A bladder that has been separated from the organism, or completely denervated, has its sphincter tonically contracted. This contraction is also observed in cases of severe injury to the spinal cord. As the bladder is filled, the internal pressure rises much more rapidly than when its innervation is intact. At first the detrusor loses its tonus, but after a time, when the intravesical pressure increases, tonic or rhythmic contractions are stimulated by filling the bladder. These contractions increase as the bladder is filled until they overcome the resistance of the sphincter and the urine is expelled. There is only an incomplete evacuation of the bladder, as the pressure soon falls below the level needed to overcome the tonically contracted sphincter. These incomplete automatic and involuntary micturitions are frequently repeated (incontinence due to overflow).

Barrington has demonstrated the existence of a center, extending from the middle of the motor nucleus of the fifth nerve to the rostral end of the hind brain. Bilateral destruction of this center permanently suppresses normal evacuation of the bladder.

There are also centers for micturition in the hypothalamus and in the cerebral cortex. The latter are responsible for the voluntary control of the bladder and urethra, the voluntary commencement and interruption of micturition, and the contraction of the external sphincter. The spinal centers are situated in the second to the fourth sacral segments; another spinal center has been located in the fourth to the sixth lumbar segments.

BIBLIOGRAPHY

Gruber, C. M., The Automatic Innervation of the Genito-urinary System, *Physiol. Rev.*, **13**, 497, 1933.

Langworthy, O., L. C. Kolb, and L. G. Lewis, "Physiology of Micturition," Baillière, London, 1940.

Newton, W. H., "Recent Advances in Physiology," Chap. IX, The Nervous Control of Micturition, J. & A. Churchill, London, 1939.

SECTION NINE

The Nervous System

CHAPTER 65

The Organization of the Nervous System

ADAPTATION TO THE environment and its variations is essential for the survival of living organisms and for the maintenance of their normal state. *Excitability* is the capacity of an organism to respond with an internal variation to changes in the environment. A *stimulus* is a modification in energy which takes place in the environment and which can evoke a response from an organism. Complex animals have specialized structures, sensitive to different kinds of stimuli, which are known as *receptors*. The excitatory state produced by a stimulus in the receptor is transmitted by *conductors* (nerves) to *centers*, where it is controlled and may be distributed by means of other conductors to the *effectors*, *i.e.*, structures that react to perform the necessary adaptation. Simpler organisms have no centers; the excitatory state is transmitted directly by conductors from the receptors to the effectors. In unicellular organisms reception, transmission, and response take place in the same cell. When a paramecium, for example, swimming in its environment by means of the regular movements of its cilia, comes up against an obstacle, the impact acts as a stimulus. An excitatory state is produced at the site of stimulation and is transmitted through the protoplasm to the cilia (the effectors) which reverse their movements. The whole organism now displaces itself in another direction owing to the adequate response that has followed the stimulus. Leukocytes and other migratory cells of vertebrates behave as isolated cells.

Multicellular organisms constitute an integrated unit. The different parts are kept together by intercellular substances in an orderly arrangement, according to laws regulating the development of form in the particular species. This static correlation makes the organism a unit in space; there are other mechanisms that assure its dynamic unity. Thus, changes in one part of the body can influence the activity of other and distant parts, provoking adaptation responses that serve to maintain the vital equilibrium of the whole animal. There are two mechanisms of dynamic correlation, the humoral and the nervous.

Cells pour the products of their activity into the circulating fluids of the body and in this way can act on other distant cells. Complex animals have an important differentiated mechanism of humoral correlation, the endocrine glands. These glands produce specific substances, known as hormones, which are secreted into the body fluids and serve to regulate the activity of certain tissues or organs. Humoral correlation is relatively slow and diffuse. The "chemical messengers" are transported at the speed of the circulating fluids, which is measured in minutes and seconds, and they act on all the cells that can respond to them.

Celenterates and animals of greater complexity have also a more rapid and precise mechanism of correlation, the nervous system, made up of cells in which excitability and the capacity to transmit the excitatory state are highly developed. The time of reaction in this system is measured in seconds and thousandths of a second. According to the circumstances the effects produced can be localized (with great precision in vertebrates) or transmitted to the whole organism. The nervous system, in those species in which it is well developed, is the principal mechanism of internal correlation and

of communication with the environment. It is the main agency by means of which the organism is able to react as a unit. This is *the integrative action of the nervous system.*

The importance of the nervous system in the maintenance of the activity of the whole organism is demonstrated by the results of its suppression. Celenterates have only a rudimentary nervous system; nevertheless anesthesia, which eliminates nervous activity, produces in them a condition of immobility and lack of activity. The nervous system is highly developed in vertebrates; therefore considerably greater effects are produced by its annulment. A decapitated frog, in which the higher coordinating nerve centers have been eliminated, remains passive in any position in which it is placed, but its muscles still have tonus and it responds to stimuli with reflex movements. If now the dorsal spinal roots are cut, no impulses coming from the periphery can reach the centers. The muscles are flaccid, because tonus depends on afferent impulses coming from the muscles and other peripheral structures, and even the strongest stimuli provoke no response. Occasionally, apparently spontaneous movements may be observed. When all the nerve centers have been destroyed the animal remains flaccid and still. Its tissues are alive, the heart beats and the blood circulates, but the animal is no longer an integrated unit, which can search for and obtain food and which can defend itself or flee from danger. The different organs continue to function with no other correlation than that of the body fluids, and soon the final disintegration of death takes place.

Disturbances in nervous functions may cause a decrease in the capacity to receive stimuli (sensory disorders) or in the coordination and execution of movement (motor disorders). Suppression of the activity of the higher coordinating centers causes loss of consciousness in those animals in which conscious phenomena can be recognized. The nervous system is the necessary instrument of intelligence and is essential for the acquisition of knowledge. The external universe (the cosmos) and the internal universe (the organism) are known through the sensorium. The normal functioning of the complex nervous mechanisms of association is indispensable for correct thinking; if these functions are perturbed, the mental faculties, *i.e.*, memory, the formation of ideas, reasoning, etc., are deranged, and an independent, free life is no lo ge r possible. There is no possibility of wisdom for man if the higher functions of his nervous system are not normally performed; knowledge cannot be acquired or integrated with activity so as to do good. Gaskell expressed its importance in a famous epigram, "*The race is not to the swift, nor to the strong, but to the wise.*" The greater the development of the nervous system, the greater is the capacity of adaptation and the surer the control of the environment. In this respect man has an eminent position among all living organisms.

Plants do not respond by rapid adaptation to changes in their environment; their conditions of life do not require it. They have no nervous system comparable to that of animals, but the excitatory state can be transmitted to organs distant from the site of stimulation and thus provoke an adequate response. Notable examples are to be found in the folding of the leaves of sensitive and carnivorous plants.

THE ELEMENTARY NERVOUS SYSTEM

The study of the general features of the nervous system of simpler animals makes it easier to understand the complicated nervous system of vertebrates.

Independent effectors. Sponges have differentiated effectors but no nervous system to initiate and coordinate their activity. On the outer surface there are many pores, covered by a membrane, which open into the subdermal spaces, from which canals lead to chambers lined by an epithelium of flagellated cells. These chambers all communicate with the central cavity. The opening of this cavity, called the osculum, and the pore canals are surrounded by elongated cells arranged to form sphincters. Currents in the water in which the sponge is submerged, several chemical compounds, and other stimuli, produce retraction or expansion of the membranes covering the dermal pores, and contraction or relaxation of the sphincters. The continuous movements of the flagella in the internal chambers keep the water circulating through the animal, from the dermal pores to the central cavity and out by the osculum. A considerable amount of water can pass through; its volume is regulated by the opening and closing of the ingoing and outgoing channels. Each one of the effectors (dermal membrane or

sphincter) acts on its own in response to stimuli that activate it. The excitatory state can be slowly transmitted for a short distance (neuroid transmission), but there is no mechanism for the simultaneous and coordinated action of several effectors. The regulation of the movements of the entrance and exit orifices results from the simultaneous response to stimuli acting on them at the same time. This simple mechanism is sufficient for the achievement of functional unity in the relatively limited and stable conditions in which these animals live.

Nearly all the effectors in vertebrates are under the influence of the nervous system, but a few respond to stimuli directly applied to them, *i.e.*, they behave as independent effectors. The muscle in the amniotic membrane has no nerves, yet it contracts and relaxes, and the heart starts beating in the embryo before it is innervated. The capacity to respond to direct stimuli remains latent in those effectors which are normally activated by the nervous system, as can be demonstrated by severing their nerves. Striated muscles are paralyzed by denervation, but even after the cut nerves have degenerated completely, they contract on application of an electric stimulus. Isolated strips of intestinal muscle, in which there are no nerve cells or fibers, contract rhythmically and show variations in tone. Fragments of heart muscle, free of nerves, continue to beat in tissue cultures long after they have been separated from the organism of which they originally formed part.

Neuroid transmission is also found in vertebrates. The movements of cilia can be modified by stimuli applied at a distance, the excitatory state being conveyed by the epithelial cells.

Systems with Receptors, Effectors, and a Nerve Net

The nervous system of celenterates (hydroids, actiniae, and medusae) consists of differentiated receptors connected by a nerve net to the effectors.

Receptors. Certain epithelial cells in the outer layers of the ectoderm and endoderm are differentiated sensory cells. Superficially they have bristlelike endings. At the internal end they emit a delicate process, which penetrates into the deeper layers and branches out, coming in contact with the processes of the nerve cells. In certain parts, the majority of the fibers follow a preferred direction, *e.g.*, in the tentacle of the sea anemone they go toward the root of the tentacle. In certain cases these receptor cells are grouped in a rudimentary sense organ, *e.g.*, the marginal bodies of the jellyfish *Aurelia*. They have an elementary specific irritability, being more sensitive to certain stimuli than to others.

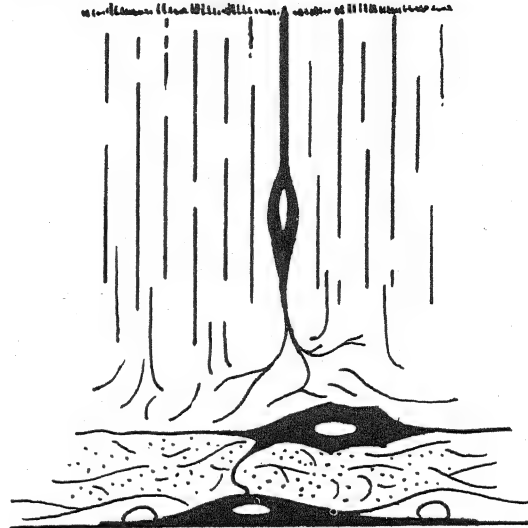



Fig. 312. Diagram of the nervous system of celenterates. Sensory or receptor cell, motor cell, and muscle or effector cell. (*Parker, G. H., "The Elementary Nervous System," J. B. Lippincott Company, Philadelphia, 1919.*)

Effectors. The ciliated cells, mucous glands, sting cells, and muscle sheets are the effectors. Some of them behave as independent effectors; thus the sting cells respond to direct stimulation, and there is no evidence that nervous excitation can cause them to discharge. Certain muscles, even though they may be stimulated through the nervous system, also respond to direct stimulation, *e.g.*, a beam of light falling on the column of an actinian will provoke localized contraction of the longitudinal parietal muscle so that the animal bends to the light. Muscles can contract to 20 per cent, or even 5 per cent, of their extended length, and they may be anywhere between full extension and maximal contraction. In some cases contraction is very slow, and the "muscle twitch" may last 5 min.; in others it is relatively rapid.

Nerve net. Lying above the muscle layer there is a plexus made up mainly of relatively large bipolar nerve cells, with two long processes or axons, which may reach 7 or 8 mm. in length. These axons have a fibrillar structure and maintain their individuality; when one axon comes

into contact with another, they do not fuse and form a syncytium. At the place where they touch there is a "synapse" (Fig. 313).[1] The axons of the nerve cells in some cases run in a preferential direction, *e.g.*, in the mesenteries of actiniae they run parallel to the muscle fibers or obliquely

FIG. 313. Parallel axon in mesentery of *Metridium* with four synaptic contacts. (*Pantin, C. F. A., Proc. Roy. Soc., London, s.B., vol. 140, p. 147, 1952.*)

from the body wall toward the retractor muscle, where they end in small expansions (end-plates, Fig. 314).

Nervous activity of celenterates. There are several types of nervous activity besides the direct response of independent effectors.

Reflex responses. Adequate stimulation will cause certain fairly rapid responses, such as contraction of the marginal sphincter or retraction of the oral disk by contraction of the mesenteric retractors. If the stimulus attains a certain threshold, a wave of excitation spreads at a definite velocity in all directions along the nerve net, followed by an absolute and relative refractory period. The response is of the "all-or-nothing" type as in vertebrates (see Chap. 66). These features are demonstrated by the following experiment: all except one of the marginal bodies are removed from a medusa; the remaining one, whichever it may be, sends out impulses which spread through the nerve net in all directions, producing waves of contraction throughout the muscle of the bell (Fig. 315). There is no polarized conduction in a definite direction, no irreciprocal spread of excitation at the points of contacts of the axon, as occurs in the synapses of vertebrates. Polarization of conduction is, however, achieved to a certain extent by the anatomical distribution of the fibers, *e.g.*, in the tentacles of actiniae the majority of sensory processes go toward the root and few toward the tip.

FIG. 314. End-plate of axon on retractor muscle of the oral disk of *Metridium*. (*Pantin, C. F. A., Proc. Roy. Soc., London, s.B., vol. 140, p. 147, 1952.*)

[1] According to Leghissa (*Pubbl. d. stazione zool. di Napoli*, **21**, 272, 1949), there is synaptic contact between the sensory and nerve cells, and between the nerve cells and muscle fibers, but the nerve net itself is syncytial. The weight of evidence is in favor of synaptic contact in the nerve net of medusae and actiniae, but the nature of the nerve net in *Hydra* is not yet definitely known. The giant-fiber system of certain worms, however, is syncytial (NICOL, J. A. C., *Quart. J. Biol.*, **23**, 291, 1948).

A single impulse usually does not provoke contraction, but it leaves a facilitating effect so that a second impulse initiates contraction. This process apparently takes place at the junction of the axon with the muscle (Pantin's neuromuscular facilitation).

Rapid reflex responses are mediated by nerve nets which conduct at relatively high speed; *e.g.*, in a metridium 10 cm. high an impulse spreads up the mesenteries in about 50 msec. This system has been called by Pantin the "through-conduction system."

Local responses. Stimulation of a tentacle or the oral disk of an actinian will awake only a local response in the neighborhood of the stimulus. Repeated application of strong stimuli (mechanical, chemical, or electrical), such as occurs in the feeding reaction, is followed first by a local response which later spreads to neighboring tentacles and may involve the whole disk. All-or-nothing impulses are generated which enable their successors to spread over wider areas activating neighboring units. The nerve net mediating these local responses seems to be distributed in units connected with each other (interneural system) so that facilitation (interneural facilitation) may take place between them and excitation spread to several units. There is evidence that the interneural and the through-conduction systems are connected, and that stimulation of one may facilitate conduction in the other.

Spontaneous activity. In widespread and local responses so far described, activity is awakened by an external agent. Celenterates also have spontaneous activity, in the sense that it has its origin within the animal. It is difficult to appreciate because of its extreme slowness; but it has been recorded by kymographic registration and by lapse-rate cinematography. This activity is periodic, the cycles lasting from several minutes to several hours. Thus actiniae expand and retract in 10 to 20-hr. cycles even when kept in complete darkness and in a stable environment. This type of activity is especially developed in certain parts of the animal, *e.g.*, the "column action system" (Jennings). Outside stimuli modify its activity usually by changing the pattern of behavior, not by provoking a definite localized or widespread response. Contraction and expansion of the different muscles are well coordinated; thus, lengthening of the column is brought about by relaxation of the longitudinal parietal muscle and waves of peristaltic contractions of the circular muscle. This sometimes involves reciprocal inhibition (see Chap. 69), *e.g.*, a local contraction of the circular muscle may temporarily inhibit the response of the parietal muscle to excitation.

R

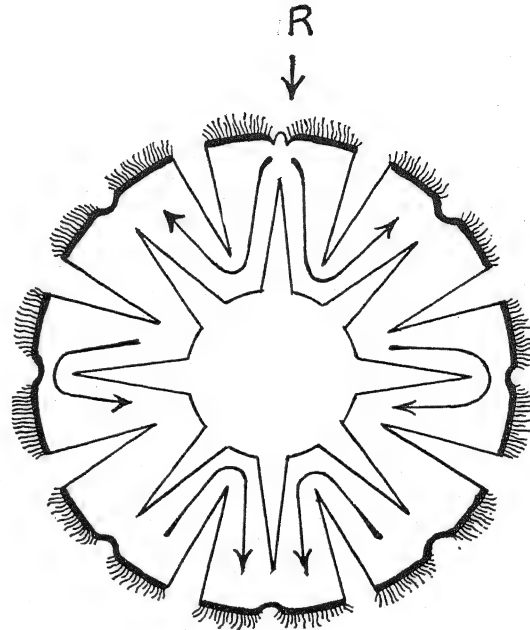



Fig. 315. Diagram of the medusa *Aurelia.* The body has been cut in several parts and all the receptor organs extirpated, except one, *R.* Stimulation of this organ provokes a wave of contraction in the muscle of the bell in the direction of the arrows. When two waves meet they annul each other. (*Parker, G. H., "The Elementary Nervous System," J. B. Lippincott Company, Philadelphia, 1919.*)

Locomotion in actiniae is carried out by this same type of slow semi-independent activity of the pedal disk and column of the animal.

The condition of the animal may modify its capacity to respond. For example, feeding or the introduction of meat juice into the esophagus results in a diminished activity of the tentacles, as if the threshold had been raised. On the contrary a starved animal may go through the whole series of feeding reactions, without any obvious outside stimulation, as if the threshold has been lowered by starvation.

Signs of fatigue may be observed when a stimulus is repeatedly applied, so that it loses its efficiency and fails to evoke a response. This is most evident in widespread and local reflex responses.

Spontaneous activity continues for a time in animals anesthetized with magnesium, in which the through-conduction system no longer responds.

The different parts have considerable autonomy; thus a tentacle which has been cut off continues to react to stimulation, and the pedal half of an actinian amputated from the rest of the animal has movements of locomotion (Parker). These observations led to the belief that celenterates should be considered as a sum of autonomous parts rather than as integrated units (Uexkull). The intact animal, however, shows coordinated activity, and "wherever autonomy is most evident it is invariably associated with the local development of specialized action systems" (Pantin).

Systems with Receptors, Effectors, and Nerve Centers

The formation of nerve centers is an outstanding feature in the nervous system of worms, arthropods, and vertebrates. Nervous structures migrate below the surface in the course of embryonic development, and consequently the nervous paths are lengthened. Afferent fibers from the receptors converge toward centers, from which final motor paths, common to many sensory areas, pass out to the effectors.

Receptors. The sensory nerve cells are not found on the surface but are deeply placed, *e.g.*, the dorsal root ganglia of the vertebrates; they come into contact with the surface only through their fibers. These peripheral fibers branch freely in the outer layer, *e.g.*, the pain fibers of the human skin (Fig. 316). More frequently the peripheral fibers of the receptor cells are surrounded by specialized epithelial cells, sometimes forming complicated structures called the sense organs. The specialized cells respond easily to one kind of stimulus; other stimuli do not provoke a response unless they are of considerable strength (specific irritability of the receptors). The afferent neurons are not stimulated directly; the stimulus acts on the specialized epithelial cells, which transmit the excitatory state to the afferent nerve fibers.

Effectors. Nearly all the effectors are controlled by the nervous system, and even though they retain a latent capacity to respond to direct stimulation, in normal circumstances their activity is regulated by nerve impulses; thus striated muscle loses all movement once it is denervated. The influence of the nervous system is not so predominant in other effectors, *e.g.*, in most of the glands and the visceral muscles, which function almost as well after complete denervation as before. There are also a few independent effectors.

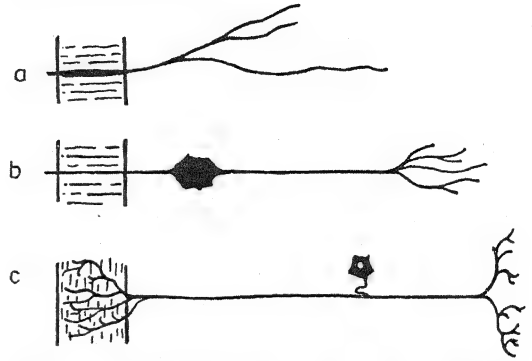



FIG. 316. Diagram of sensory cells. *a*, in celenterates; *b*, in worms; *c*, in vertebrates. (*Parker, G. H., "The Elementary Nervous System," J. B. Lippincott Company, Philadelphia, 1919.*)

The motor nerve cells also migrate into the deeper layers. Their polarization is complete; they are stimulated either directly on the cell body or by one of the fibers (dendrites), and they transmit the excitatory state exclusively by one special fiber, the axon (Fig. 317). This contrasts with what occurs in celenterates, where the excitatory state can spread in all directions.

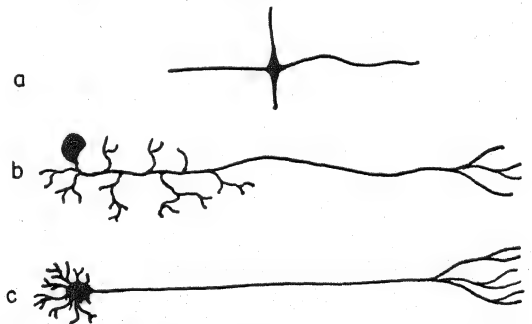



FIG. 317. Diagram of the evolution of motor cells. *a*, in celenterates; *b*, in worms; *c*, in vertebrates. (*Parker, G. H., "The Elementary Nervous System," J. B. Lippincott Company, Philadelphia, 1919.*)

Nerve centers. The bodies of the cells are grouped together in deeply placed structures such as the dorsal root ganglia and the motor nuclei of the ventral horns of the spinal cord. Exceptionally they can be found in the receptor (retina) or the effector (intestines). The connec-

tion between neurons is established by contact, without continuity between the cells. This connection is called the *synapse*[1] (Greek συναπσις, clasping). The shortest paths are made up of only two nerve cells or neurons; one is connected with the receptor (afferent neuron) and the other with the effector (efferent neuron), and they are united by a synapse (Fig. 318*a*). The majority of paths are made up of many neurons; several connector or internuncial neurons are placed between the afferent and efferent neurons (Fig. 318*b*). The number of connector neurons increases with the complexity of the nervous system. The cell bodies of the connector neurons are also grouped in centers which regulate the activity of the lower motor centers. The paths connected with the cephalic receptors acquire great importance in vertebrates; the internuncial neurons of these paths form highly developed centers which are connected with many other centers. This process, known as "encephalization," progresses as the nervous system becomes more complicated. These centers control all the others in the integration of complicated reactions, while the lower centers coordinate only rudimentary sensibility and simple movements. The receptive capacity of the sensory organs and the variety and precision of movement increase with the development of these higher centers.

FIG. 318. Reflex arcs. *a*, bineuronal reflex arc; *R*, receptor; *SN*, sensory neuron; *MN*, motor neuron; *Sy*, synapse; *E*, effector. *b*, multineuron arc; one or more internuncial (*IN*) or connector neurons are interposed between the sensory and motor neurons.

In this type of nervous system the impulse is transmitted in a determined direction (Sherrington's law of forward direction of conduction); there is no diffuse radiation of the excitatory state, such as is seen in the nerve net of celenterates. This outstanding feature is not due to any specific property of neurons in these species, but to the manner in which they are connected with each other. The excitatory state diffuses through the whole neuron, cell body and fibers, but it is not transmitted to all the cells connected with the excited neuron—only to those with which the axon makes a synapse. In celenterates there is either protoplasmic continuity between all nerve cells, or else the membranes that separate them can transmit the excitatory state with equal ease from one cell to another, whatever the direction from which it arrives at the connecting surface. Vertebrate neurons, on the other hand, are anatomic units, surrounded by a membrane which conditions the irreciprocal conduction of the excitatory state and thus causes complete polarization of the neuron in the intact, normal nervous system.

Two different processes can take place: the *central excitatory state*, which transmits excitation from one neuron to the other, and the *central inhibitory state*, which annuls the former and puts a limit to its diffusion (see Chap. 69).

[1] Sherrington introduced this term. For further details see FULTON, J. F., "Physiology of the Nervous System," 3d ed., Oxford, New York, 1949.

THE NEURON

The nerve cell, or neuron, is the morphologic unit of the nervous system. It consists of a nucleus, surrounded by a protoplasmic mass, the *perikaryon*, and protoplasmic processes. The nucleus is usually large, spherical, and poor in chromatin, and it has a nucleolus. The protoplasm contains mitochondria, a Golgi apparatus, and two kinds of pigment. One of these, melanin, is found in great quantities in the cells of the locus coeruleus and the substantia nigra; the other is a yellow pigment, which increases in old age. There is also a substance which cannot be seen in live cells but which in fixed cells precipitates into granules, known as the Nissl bodies. These bodies are colored darkly by basic stains, such as methylene blue; they are found in the perikaryon and the roots of the dendrites. This substance probably plays a part in the metabolism of the cell, as it tends to disappear when the cell is injured or fatigued, a process called chromatolysis. There are numerous anastomosing fibrils in the protoplasm which pass into the dendrites and axons, running parallel to the main axis of the fiber.

Neuron processes are of two kinds: (*a*) the *dendrites;* (*b*) the *axons.*

The dendrites branch out in the proximity of the cell to which they belong. Most neurons have several dendrites, but some, among which are the sensory cells in the spinal root ganglia, apparently have none. In the course of embryonic development the dendrites of these cells take on the appearance of axons, although genetically and functionally they are dendrites, conducting impulses toward the cell body.

There is only one axon or axis cylinder in each neuron, although a few have two. These processes arise from a small eminence in the cell body, usually free from Nissl bodies, known as the "axon hillock," and they send off branches at right angles to the principal fiber. Some neurons have very long axons, which can be over 1 m. in length; they are called Golgi's type I neurons. Others have short axons, which send off branches near the cell body; they are Golgi's type II neurons. Axons branch off into numerous very thin fibers, which end in a swelling, known as the terminal bulb, *bouton*, end-knob, synaptic knob, or end-foot. The synaptic knobs come into contact with the dendrites or the perikaryon of another neuron. The site of contact is the synapse. Each neuron receives many of these knobs. On the large pyramid cells of the frontal cortex there are several thousand of them, and about one million on the Purkinje cells of the cerebellar cortex; the motor neurons of the human spinal cord have up to 23 per 100 sq. μ of cell surface (Fig. 319), and the sensory neurons up to 14 per 100 sq. μ. Approximately 38 per cent of the cell surface of the spinal motor neurons of the cat is covered by terminal bulbs. Large club-shaped endings have been described in certain small cells of the tegmentum; the area of contact is approximately 50 sq. μ instead of the 1 to 2 sq. μ of the ordinary synapse. This club ending covers about 10 per cent of the surface of the cell with which it comes into contact.[1] These facts are of importance in the interpretation of synaptic transmission (see Chap. 69). The dendrites and axon endings sometimes form a thick network around the neuron, which is called the neuropile.

The perikaryon and the dendrites are the receiving part of the neuron. This part has been called the *soma*[2] to differentiate it from the axon, which transmits the excitatory state or nerve impulse to other cells.

Nerves. These structures are the peripheral projection of the neuron processes. They are made up of fibers, *i.e.*, axons and their sheaths.[3] The diameter of the axons varies from 1 and 2 μ in the finer ones to 18 and 20 μ in the larger fibers. Many functional properties of axons, such as their excitability, refractory state, and speed of conduction, are related to the diameter.

Under the phase contrast microscope, axons appear to be made up of a homogeneous, doubly refracting substance, within which there are numerous dark bodies similar to mitochondrias. There are no signs of fibrillar structure.

The ultrastructure of the axon has been studied by means of the electron microscope in isolated giant axons of invertebrates; in vertebrate nerves after dissociating their fibers by the action of ultrasonic waves; in very thin sections; in tissue cultures and in axons extruded from their sheaths. The axon appears to be a

[1] BODIAN, D., *Physiol. Rev.*, **22**, 146, 1942.

[2] LORENTE DE NÓ, R., and H. T. GRAHAM, *Am. J. Physiol.*, **121**, 388, 1938.

[3] Some of these are axons in the strict sense of the term, *i.e.*, they conduct impulses away from the perikaryon. Others are genetically and functionally dendrites, which in the course of development have acquired the aspect of axons; they conduct impulses toward the perikaryon.

large bundle of tightly packed fibrils placed in the longitudinal axis. The diameter of each fibril is 100 to 400 A (1 A = 0.1 mμ). Frequently the fibrils show a higher optical density along the edges, taking on the aspect of tubes (neurotubules). After the axon has been cut off from the perikaryon, the fibrils disintegrate very quickly, simultaneously with the fall in action potential. Between the fibrils there are many bodies of high optical density. The largest of these are rod-shaped, measuring 250 to 900 mμ; they have the appearance of mitochondria.[1]

The outer layer of the axon, the axilemma, acts as a membrane; inductance, resistance, capacity, and other electrical properties of this membrane have been measured in various physiologic conditions, in several species. It has a selective permeability; therefore the axoplasm differs from the surrounding media, *e.g.*, the concentration of K^+ is higher and that of Na^+ and Cl^- lower than in extracellular fluid.

In the course of development, when the diameter of the axon is between 1 and 2 μ, it is surrounded by a sheath of concentric layers of protids and lipids. This is the myelin sheath, which is interrupted at regular intervals by constrictions called the nodes of Ranvier. Studies with the electron microscope[2] have shown that the myelin sheath is made up of concentric lamellae 80 A thick; there are about 250 of these in a 2-μ sheath, separated by layers of water and protein. X-ray diffraction studies indicate that unimolecular layers of protein alternate with bimolecular layers of lipid. The thicker fibers have a relatively thinner myelin sheath than finer fibers. Myelin has a high electric specific resistance, about ten million times that of Ringer's fluid, and it acts as an insulator.

A thin protoplasmic layer in which there are nuclei surrounds the myelin sheath; it is clearly visible only around the nuclei and at the nodes

0 10μ 20μ

a

0 5μ 10μ 15μ

b

FIG. 319. Synapses in motor neurons. *a*, neuron in gray matter of the spinal cord of the cat, showing many degenerating "synaptic knobs" in contact with the cell surface; *b*, detail of one "knob." (*Hoff, in Creed et al., "Reflex Activity of the Spinal Cord," Oxford, New York, 1942.*)

[1] DE ROBERTIS, E., and F. O. SCHMITT, *J. Cell. & Comp. Physiol.*, **31**, 1, 1948; **32**, 45, 1948; *J. Exper. Med.*, **90**, 283, 1949; DE ROBERTIS, E., *J. Exper. Med.*, **90**, 291, 1949; SCHMITT, F. O., *J. Exper. Zool.*, **113**, 499, 1950; SCHMITT, F. O., and B. B. GREEN, *J. Exper. Med.*, **91**, 499, 1950; FERNÁNDEZ-MORÁN, H., *Exper. Cell. Res.*, **1**, 309, 1950; **3**, 282, 1952; DE ROBERTIS, E., and C. M. FRANCHI, *Arch. Soc. de Biol. de Montevideo*, **17**, 112, 1950; DE ROBERTIS, E., C. M. FRANCHI, and J. R. SOTELO, *Publ. Inst. Invest. Ciencias Bio. Montevideo*, **1**, 173, 1951; DE ROBERTIS, E., and J. R. SOTELO, *Exper. Cell. Res.*, **3**, 433, 1952; DE ROBERTIS, E., and C. M. FRANCHI, *J. Exper. Med.*, **98**, 269, 1953.

[2] FERNÁNDEZ-MORÁN, *loc. cit.*

of Ranvier. This layer is formed by the Schwann cells. A thin membrane, the neurilemma, covers the whole fiber; it comes into contact with the axon at the nodes of Ranvier, and it is only at these nodes that the axon sends off branches. The neurilemma is surrounded by another sheath made up of reticular fibers with a few collagenous fibers (Fig. 320*a*). The fibers of a nerve are kept together by a nucleated fine connective-tissue structure, the endoneurium, and the whole nerve is surrounded by a connective-tissue sheath, the epineurium.

Some nerve fibers do not have a myelin sheath; they are called Remak's fibers. The axon is of small diameter and is surrounded by a thin layer of lipids under the Schwann cells and the neurilemma (Fig. 320*b*). Most of the postganglionic fibers of the sympathetic are unmyelinated. This type of fiber is also found in sensory nerves in a proportion varying from 1:1 to 1:4 with respect to myelinated fibers. Section of the dorsal spinal roots, distal to the ganglia, causes degeneration of all the myelinated fibers and 80 to 90 per cent of unmyelinated fibers in peripheral nerves. The latter conduct impulses originated by painful stimuli.

There are four types of nerve fibers: (*a*) naked axons, found in the gray matter of the nerve centers and the peripheral endings of axons; (*b*) axons surrounded by a myelin sheath, but without the neurilemma, found in the white matter of nerve centers; (*c*) axons with myelin and neurilemma sheaths, which are the principal constituents of most peripheral nerves; (*d*) axons with no myelin, but only the Schwann cells and neurilemma (Remak's fibers), which are the postganglionic fibers of most visceral nerves and an important part of the sensory fibers of peripheral nerves.

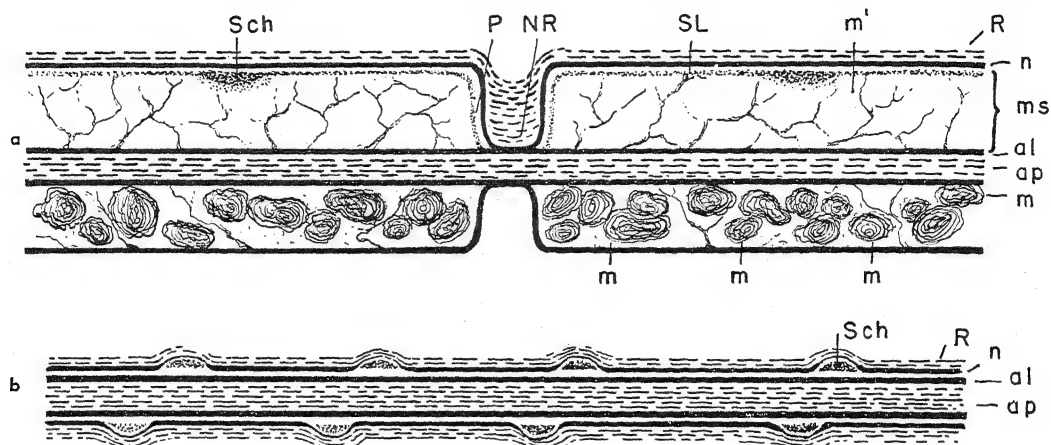



Fig. 320. Myelinated and unmyelinated fibers. *a*, diagram of myelinated fiber; *ap*, axoplasm; *al*, axilemma; *ms*, myelin sheath; *m*, myelin; *m'*, space left by dissolved myelin; *n*, neurilemma; *SL*, Schmidt-Lantermann clefts; *Sch*, nuclei of Schwann cells; *P*, protoplasm of Schwann cells; *NR*, node of Ranvier; *R*, Retzius membrane. *b*, Remak's (unmyelinated) fiber.

Neuroglia. There is no connective tissue in the nerve centers, except around the blood vessels; neuroglia is the supporting tissue. Neuroglia proper includes several types of cell: (*a*) *protoplasmic astrocytes*, with multiple branching processes, which are the mossy cells in the gray matter of the spinal cord and brain; (*b*) *fibrous astrocytes*, with long unbranched fibers, which are the spider cells of the white matter; (*c*) *oligodendroglia*, small cells with very thin processes, few in number and rarely branching, which are found in close proximity to neurons and nerve fibers and to which metabolic functions have been attributed. In types (*a*) and (*b*) some of the processes are attached by "end-knobs" to the blood vessels. The surface of the brain and spinal cord is covered by the piaglial membrane, on the inner aspect of which there are numerous astrocytes. This membrane sends sheaths around the blood vessels that penetrate into the nerve centers

The ependymal cells are also supporting cells; they line the ventricles of the brain and the central canal of the spinal cord. These cells have a long, slender process, which goes into the

nerve tissue. In the embryo they also have numerous cilia on the free surface of the cell, but most of these are lost in the adult.

The white and gray matter of the brain and cord have small cells of mesodermal origin, which constitute the *microglia.* They have little protoplasm, two or three thorny, branching processes, and a small nucleus. They are phagocytic cells and play an important part in the removal of dead tissue.

The Trophic Function of the Neuron

The neuron is a trophic unit; the nucleus and the perikaryon play an indispensable part in the maintenance of the anatomic and functional integrity of the whole cell. When a process is cut, the part separated from the perikaryon disintegrates while the cell reacts and repairs the damage, provided the injury is not such that it kills the whole cell.

Degeneration of nerve fibers. After a nerve is cut or crushed, the part of the fiber separated from the perikaryon suffers profound changes in the axon and sheaths, which are known as "wallerian degeneration," after Waller, who was the first to describe them. The axon shortens and folds, and later is split into fragments. The fine submicroscopic tubes visible under the electronic microscope rapidly disintegrate, and the orderly distribution of molecules and micellae is lost.[1] Myelin also is split up into fragments. Phagocytic mesenchymatic cells invade the neural tube a few hours after the nerve has been cut and have been found up to 80 days after nerve section. The neurilemma sheath and endoneurium thicken and coalesce, the Schwann-cell nuclei multiply, the protoplasm increases, and the whole cell lengthens, taking on a fibrous aspect. Gradually the Schwann cells fill all the neural tube. This tube narrows down so that around the fiftieth day it is reduced to about half the original diameter.

Chemical changes occur in the course of degeneration and regeneration of nerve. The cut nerve swells and its wet weight increases. Neutral tat, found mainly in the connective-tissue sheath, decreases rapidly. Chemical disintegration of the myelin sheath does not begin until about 8 days arter the nerve has been cut. The myelin lipids (free cholesterol, cerebroside, and sphyngomyelin) diminish;[2] cholesterol is in part bound to the fatty acids set free by hydrolysis, and cholesterol esters appear; kephalin and lecithin, which are axon constituents, also decrease, the former more rapidly and the latter more slowly than the myelin lipids. Phosphoprotein and a lipoprotein complex which forms part of "neurokeratin" also diminish.

Nucleic acids increase coincidently with cell proliferation (Schwann cells, fibrocytes, macrophages). The curve representing changes of desoxypentosenucleic acid (DNA) is similar to the curve representing changes in the total number of cell nuclei. Pentosenucleic acid (PNA), an index of total mass of cytoplasm, increases more than DNA. The PNA:DNA ratio increases at the same time that an increase in Schwann-cell cytoplasm can be demonstrated.[1]

Acetylcholine and the enzyme system for its hydrolysis and synthesis (cholinesterase, cholinacetylase) diminish.[2] Other enzyme systems also undergo marked changes.[3] Acid phosphomonoesterase increases at the same time as DNA. Alkaline phosphomonoesterase diminishes. 5-Nucleotidase, active on substrates such as adenosine-5-phosphate and inosine-5-phosphate, increases.

During regeneration myelin lipids increase, spreading toward the periphery as myelin is built up, but they remain below normal concentration for a long time, perhaps permanently. Kephalin and lecithin also increase, but the latter does not return to the normal level; neither does neurokeratin. The concentration of nucleic acids remains high. The regenerated nerve differs chemically from normal nerve, as it differs by some of its physiologic properties (see "Regeneration of nerve fibers").

Degeneration of mammalian nerves is complete in about four days, but the different types of fibers do not degenerate at the same rate. Many years ago histological observation showed that certain sensory nerves degenerated more rapidly than motor nerves,[4] and recently[5]

[1] YOUNG, J. Z., *Nature,* **153,** 333. 1944.

[2] JOHNSON, A. C., A. R. MCNABB, and R. J. ROSSITER, *Biochem. J.,* **45,** 500, 1949; BURT, N. S., *et al., Biochem J.,* **47,** 318, 1950.

[1] LOGAN, J. E., *et al., Biochem. J.,* **51,** 482, 1952.

[2] FELDBERG, W., *J. Physiol.,* **101,** 432, 1943; VON MURALT, A., and G. VON SCHULTHESS, *Helvet. physiol. et pharmacol. acta.,* 2, 435, 1944; NACHMANSOHN, D., *et al., J. Biol. Chem.,* **163,** 475, 1946; SAWYER, C. H., *Am. J. Physiol.,* **146,** 246, 1946.

[3] HOLLINGER, D. M., *et al., Biochem. J.,* **52,** 652, 1952.

[4] MONCKEBERG, G., and A. BETHE, *Arch. f. mikr. Anat.,* **54,** 135, 1899.

[5] GUTMANN, E., and J. HOLUBAR, *Nature,* **163,** 328, 1949.

registration of action potentials (see Chap. 66) has confirmed the fact that sensory fibers from the skin degenerate earlier than afferent and efferent fibers in motor nerves. Rosenblueth and Dempsey[1] have seen that thick fibers with a rapid conduction velocity cease to conduct nerve impulses before thin, slow-conducting fibers. There is some discussion as to whether degeneration progresses centrifugally (Parker, Rosenblueth, *et al.*) or centripetally, *i.e.*, whether the fibers cease to conduct earlier in the proximal or the distal end. Erlanger and Schoepfle[2] maintain that the evidence is consistent with abrupt failure of conduction at random loci increasing in frequency toward the periphery. They suggest that differences in the blood supply of the degenerating fibers may account for the differences observed.

The proximal end also shows alterations. The axon swells, and for a few millimeters of its length above the cut it is destroyed. Myelin is broken up and ingested by phagocytes, and the Schwann cells multiply. Sometimes the axon degenerates in its whole length up to the perikaryon. The cell body swells. The nucleus is displaced to an eccentric position. The Nissl bodies disappear from the base of the dendrites on the first day, and later from the rest of the perikaryon. This phenomenon is known as chromatolysis. It is more marked when the fiber is cut near the cell body and may be almost imperceptible when only the end of the fiber is injured. With time, the Nissl bodies reappear.

Gudden's atrophy. Retrograde cell degeneration is more severe in certain cases. Gudden saw that a destructive injury of the cerebral cortex made at birth was followed by degeneration of those nuclei in the thalamus which gave rise to the fibers destroyed in the cortex.

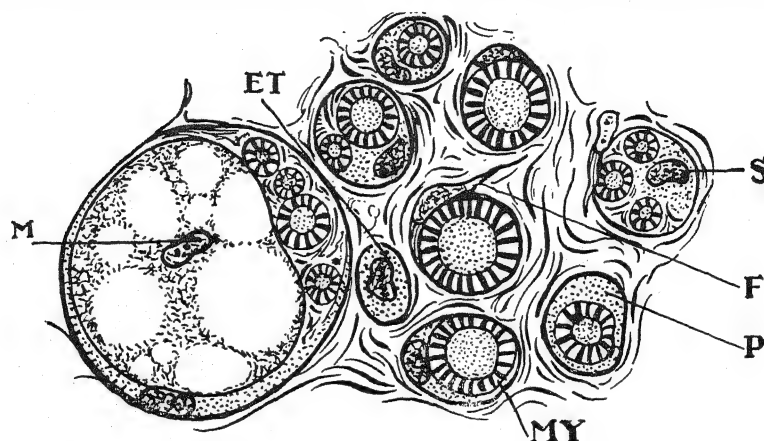



Fig. 321. Transverse section of the peripheral stump of rabbit's nerve, cut 150 days previously and left unsutured. The Schwann tubes contain one or more fibers, surrounded by a myelin sheath *MY*, and Schwann cells, with protoplasm *P* and nuclei *S*. *ET*, Schwann tube into which no fiber has penetrated; *M*, macrophage in a tube that contains several fibers in the process of regeneration; *F*, fibroblast. (*Young, J. Z., Physiol. Rev., vol. 22, p. 318, 1942.*)

Transneuronal degeneration. Destruction of neurons can provoke degeneration of the neurons that are usually stimulated by the destroyed nerve cells. A restricted lesion in the retina causes atrophy in the neurons of the geniculate body on which the axons of the destroyed retinal neurons end. This phenomenon occurs when injury has damaged the neurons that send the greater part of the impulses received by a nucleus; it can be considered as a case of "isolation atrophy."

Regeneration of nerve fibers. (Fig. 321.) The process of regeneration of an injured nerve fiber is accomplished by the joint activity of the perikaryon, the axon, and its sheaths. The Schwann cells play a very important part; they multiply and migrate from both ends, and can form a bridge up to 3 mm. in length. They are accompanied by the fibroblasts of the endoneurium, and the fibers formed by these cells give a certain solidity to the bridge. As soon as retrograde degeneration has ceased in the proximal end of the axon, a great number of

[1] Rosenblueth, A., and E. W. Dempsey, *Am. J. Physiol.*, **128**, 19, 1939.

[2] Erlanger, J., and G. M. Schoepfle, *Am. J. Physiol.*, **147**, 550, 1946.

slender processes, sometimes as many as 100, are put out from the axon and grow in all directions. Their ends take on a bulbous aspect wherever they come up against an obstacle, such as a fibrous plane, that hinders their growth. These fibers form a thick network and give rise to nerve tumors (neuromas), which can be very painful.

When an axonic process comes into contact with a Schwann cell, this cell acts as a support along which the axon grows until it reaches the distal end of the cut nerve, eventually penetrating into the neural tube. The axon grows very slowly in the midst of scar tissue, only 0.25 mm. per day, but once it has entered a neural tube it grows at a greater speed, as much as 3 to 4 mm. daily.

Axons not only grow in length but also increase in diameter from the perikaryon down. Recovery of the physiologic properties progresses in linear relation to this increase in diameter. A small action potential (see Chap. 66), conducted at a speed of less than 1 m. per sec., can be registered in the distal end of a cut and sutured nerve of a cat 17 days after operation. With time the size, velocity, and complexity of the action potential increases, but even a year after, recovery of axon size is only 80 per cent of the normal.[1] Connection with an end organ is of great importance for the increase in diameter; when a fiber connects with an end organ it develops much faster than others in the same nerve.[2]

The importance of peripheral connections is well illustrated by the following experiment: The nerve to the middle head of the gastrocnemius of the rabbit, which has about 350 fibers, was crushed on both sides. On one side it was also cut lower down so that the regenerated fibers were unable to reach the periphery. Fiber size above and below the crushed points was measured 100 days after operation. Below the crushed point on the disconnected side there were 2,000 fibers; the largest were 14 μ (in normal nerve the largest fibers were 20 μ), and the majority were much smaller. On the side where peripheral connections had been established there were nearly 600 fibers, the largest being of nearly normal size (18 μ), and size distribution was similar to that in the normal nerve. Moreover, fibers above the crushed point on the disconnected side had shrunk considerably in diameter.[1] Fibers are, therefore, influenced not only by their central connections but also by their peripheral ones—a condition that has been called by Young one of "double dependence."

The myelin sheath is also regenerated when the axon belongs to a myelinated fiber. The nature of the neural tube does not condition the formation of a myelin sheath. Thus the axon of a somatic neuron that penetrates a neural tube of a degenerated sympathetic fiber is eventually surrounded by a myelin sheath, but on the other hand if the axon of a sympathetic neuron enters a tube that has belonged to a somatic fiber, a myelin sheath is not regenerated. Nevertheless the size of the neural tube limits the growth in diameter of the axon; therefore the recovery of its physiologic properties.[2]

The anatomic and functional restoration of a nerve depends on the penetration of its axons into the distal neural tubes. Therefore, to obtain good results rapidly, this penetration should be helped. When a nerve is severed by compression, the axons are destroyed without interrupting the continuity of the nerve; regeneration then takes place at a speedier rate, and the functional deficit is finally less marked than if the nerve had been cut and its ends separated. The suture of a cut nerve with approximation of the ends is of great value; it should be done with material that does not provoke an inflammatory reaction. Stretches of nerve kept frozen and dried, and hydrated at the time of use, have been employed to fill wide gaps due to extensive destruction of a nerve.[3] The importance of the fiber pattern of the funiculi used in grafting has been pointed out.[4] The graft should possess mainly large myelinated fibers, with as little branching and plexus formation as possible. Segments of cutaneous nerves such as the superficial radial and sural nerves are very suitable for this purpose. A serious obstacle encountered in nerve regeneration is the tendency of the free axon

[1] Berry, C. M., H. Grundfest, and J. Hinsey, *J. Neurophysiol.*, **7**, 103, 1944.

[2] Sanders, F. R., and J. Z. Young, *Nature*, **155**, 237, 1945.

[1] Aitken, J. T., M. Sharman, and J Z. Young, *J. Anat.*, **81**, 1, 1947.

[2] Sanders, F. R., and J. Z. Young, *J. Physiol.*, **103**, 119, 1944.

[3] Weiss, P., *Arch. Surg.*, **46**, 525, 1943; *J. Neurosurg.*, **1**, 219, 1944.

[4] Sunderland, S., and L. J. Ray, *Brain*, **70**, 75, 1947.

tips to branch out and form neuromas. This can be prevented by encasing and joining the cut ends in a cuff, made by a fresh or adequately preserved artery or by tantalum foil. The nerve should fit snugly into the guiding tube, but it should not be tightly packed; otherwise the development in diameter of the axons will be limited, with unfavorable results in the recovery of physiologic properties. The suture of a cut nerve is successful even when made 1 month after the injury. It should not be delayed too long, however, because as time passes, the reproductive activity of the Schwann cells diminishes, the neural tubes become narrower, and the terminal organs, such as the end-plates, are atrophied.

Functional restoration of the cut nerve depends on (*a*) the entrance of a sufficient number of axons into neural tubes of appropriate diameter; (*b*) the establishment of adequate connections between fibers and end-organs, *e.g.*, between motor fibers and muscle end-plates; (*c*) the condition of the effector, which may have suffered irreversible atrophy if it has remained without functioning for too long a period.

The neurons in the nerve centers have lost all reproductive power; if they are destroyed they cannot be replaced. Spinal-cord transection in man is not followed by regeneration of the spinal tracts, even when done with surgical care and asepsis, as in operations performed to relieve pain. The absence of Schwann cells, which play so important a part in peripheral nerve fiber regeneration, and the complexity of central connections may account for some of the difficulty in the regeneration of severed or injured tracts in the nerve centers.

BIBLIOGRAPHY

Pantin, C. F. A., The Elementary Nervous System, *Proc. Roy. Soc., London, s.B.*, **140**, 147, 1952.

Parker, G. H., "The Elementary Nervous System," Lippincott, Philadelphia, 1919.

Ramòn y Cajal, S., "Estudios sobre la Degeneración y Regeneración del Sistema Nervioso," 2 vols., Madrid, 1913.

Ramón y Cajal, S., Las Pruebas Objetivas de la Unidad Anatómica de las Células Nerviosas, *Arch. de neurobiol. psicol.*, **13**, 217 and 579, 1933.

Ranson, S. W., "The Anatomy of the Nervous System," 7th ed., Saunders, Philadelphia, 1943.

Young, J. Z., The Functional Repair of Nervous Tissue, *Physiol. Rev.*, **22**, 318 and 374, 1942.

Young, J. Z., Growth and Plasticity in the Nervous System, *Proc. Roy. Soc., London, s.B.*, **139**, 18, 1951.

Excitability.

Electrical Phenomena in Tissues

EXCITABILITY IS THE capacity of the organism to respond by an internal modification (adaptation) to a change in the energy potential of the environment, *i.e.*, to a stimulus. Excitability is a property of all living cells, but it is highly developed in some tissues (*e.g.*, nerve and muscle) and less obvious in others (*e.g.*, bone, cartilage, and adipose tissue). Notwithstanding the considerable differences observed in the excitability of different cells and in one and the same type of cell in different species, there are laws of excitability common to all living organisms.

THE GENERAL LAWS OF EXCITABILITY

Specific irritability. Energy in all forms, mechanical, chemical, electrical, heat, light, sound, etc., can act as a stimulus. This does not mean that all changes in energy potential excite all cells; certain conditions, particular to the physiologic state of the cell, must be satisfied for a given stimulus to evoke a response. A fundamental condition is that the form of energy must be suitable for the receptor. Each cell has a stimulus, or group of stimuli, to which it responds with greater ease, *i.e.*, the response is provoked with a smaller quantity of energy than that needed with other forms of energy. This is known as the "specific irritability" of cells. For example, mechanical deformation of the skin stimulates certain cutaneous receptors, others are sensitive to changes in the temperature of the skin, and in the retina there are cells sensitive to light. It is also possible to excite a cell with a nonspecific stimulus. Thus a sensation of light can result from a blow (mechanical energy) on the eyeball. Similarly in physiologic conditions striated muscle fibers are stimulated exclusively by nerve impulses, but they can also be stimulated by an electric shock, a sharp blow, a sudden rise in temperature, and several chemical substances. Certain variations in energy potential apparently do not act as stimuli; *e.g.*, ultraviolet and infrared rays are not visible because they do not stimulate the retina, but they have been shown to evoke responses from other cells. It is difficult to state definitely that a given change in energy potential has no influence on an organism, because sometimes its effects are not easily recognized or may take some time to appear.

Electricity is the form of energy most commonly used for experimental purposes, because it is easily handled, can be accurately measured, and does not produce irreversible changes in the tissues when properly applied. Other forms of energy such as light, sound, mechanical energy, etc., must be used when studying special receptors sensitive to these stimuli.

Conditions of efficiency in a stimulus. Excitability can be measured, and its changes followed, by estimating the conditions that a stimulus must fulfill to evoke a response. Thus, if the excitability of a tissue diminishes, it is necessary to increase the strength of a previously adequate stimulus to evoke the same response.

Three parameters must be considered in a stimulus: (*a*) the intensity or strength; (*b*) the time in which the change of energy potential takes place, *i.e.*, the rate of change; (*c*) the time

during which this variation must remain at or above a certain level.

Strength of stimulus. The threshold. Whatever be the nature of the stimulus, it must be of a certain minimum strength to evoke a response. When the strength is just sufficient to provoke the minimum response, it is called the *threshold* or *liminal* stimulus. If it is weaker and does not provoke a response, it is called *subliminal.* A *maximal* stimulus evokes the greatest response possible, and a *supramaximal* stimulus is one of even greater strength. A *submaximal* stimulus provokes a response intermediate between the smallest and the greatest. The foregoing statements are valid for structures made up by several excitatory units; each unit, *e.g.*, a single nerve or muscle fiber, responds maximally to the threshold stimulus, *i.e.*, further increase in the strength of the stimulus does not increase the response (see "The all-or-nothing" law, page 776). When excitability increases, a previously subliminal stimulus may attain or surpass the threshold; on the contrary, when excitability decreases, a previously efficient stimulus may become subliminal.

The strength or intensity of a stimulus is given by the degree of variation in energy potential, *e.g.*, in an electric stimulus, it depends on the voltage; in stimulation by heat, on the rise or fall in temperature; etc. The quantity of energy should really be taken into account instead of considering only the rise or fall in potential. For example, in a mechanical stimulus (considered as the energy developed by the fall of a body) the height of the fall and the weight of the mass falling are the components that determine the strength of the stimulus. The efficiency of an electric stimulus depends on the density of current, *i.e.*, intensity of current i, expressed in amperes, in relation to the surface on which it acts; $i \times$ sq. cm. It is usual to express the intensity of an electric stimulus by the voltage; this is justified if there is no change in the stimulated surface or in the impedance of the circuit, as by Ohm's law, $i = V/R$ (V = voltage, R = resistance). Hence, when R is constant i varies as, and in proportion to, V. In the discussion that follows, the intensity or strength of an electric stimulus will be expressed in volts.

FIG. 322. Diagram of strength, slope, and duration of stimulating currents. *a*, the current is established and stopped instantaneously, and lasts a certain time at a given voltage; the threshold is reached and stimulation takes place; *b*, a current of the same type and strength, but lasting a shorter time, does not stimulate; *c*, the current lasts the same time as in *b*, but is of a higher voltage, therefore it acts as a stimulus; *d*, the current rises gradually, therefore it must rise to a higher voltage in order to reach the threshold; *e*, the current rises at a slower rate than in *d*, therefore it must rise to an even higher voltage to attain the threshold; *f*, the slope of the current is such that it can never reach the threshold.

The variable phase in the evolution of energy is of paramount importance; the constant phase is important only during a short initial period (see "The duration of the stimulus," page 773). This can be demonstrated by stimulating a muscle with a galvanic current. There is a contraction at make and another at break, but while the current flows at a constant potential there is no contraction. The rise in potential at make and its fall at break act as stimuli.

Excitation takes place at the cathode at make and at the anode at break. Anode-break excitation is often absent, *e.g.*, in freshly dissected muscle and in frog nerves with normal circulation. Conditions that depress a preparation and lower the membrane potential (see page 779) are favorable for the appearance of anode-break excitation which may evoke a repetitive response.

The rate of change in potential. Du Bois Reymond's law. Sudden variations in energy potential are more efficacious as stimuli than gradual changes. As the rate of change diminishes the rise or fall in potential must be increased in order to attain the threshold; if it is below a certain value no response is evoked however great the variation in potential (Fig. 322). Du Bois Reymond expressed this fact in the form of a law (Du Bois Reymond's law of excitation): Excitation is a function of the differential coefficient of the density of current c with respect to time t; dc/dt.

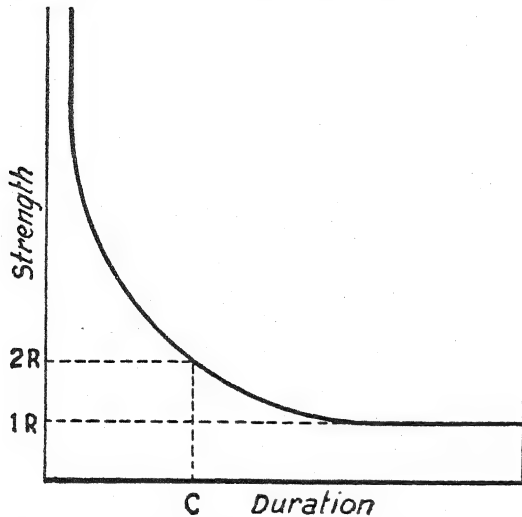



FIG. 323. Chronaxie. Strength-duration curve. $1R$, rheobase; $2R$, twice the voltage of the rheobase; C, chronaxie. (*Lapicque.*)

The duration of the stimulus. Chronaxie. A stimulus must remain at or above the threshold for a certain time, otherwise it does not evoke a response. This fact was demonstrated by the application of stimuli of very short duration, obtained by means of the ballistic rheotome, or Lucas's pendulum and spring rheotomes, or condenser discharges. At present square pulses of variable strength and duration electronically generated are used.

The strength of a stimulus and the time that it must last to excite are related in a way which is characteristic for each tissue and which can serve to define its excitability. Lapicque has given the name "chronaxie" (Greek χρονος, time, and αξια, value) to this relation, called by Waller "characteristic of excitability," and "excitation time" by Lucas. By plotting the smallest strength (voltage) of a stimulus against its duration, strength-duration curves, also called "chronaxie curves," are obtained (Fig. 323). These curves have approximately the same slope for very different tissues if the time scale is changed; this should be in seconds for "slow" tissues and in fractions of milliseconds for "fast" ones. Chronaxie curves have been obtained not only for plain and striated muscle and nerves of vertebrates, but also for invertebrates, protozoa (*Vorticella*), algae (*Spirogyra*), and sensitive plants. Chronaxie is therefore a characteristic of excitability of all cells (Fig. 324).

The excitability of a tissue can be measured by determining one point on the strength-duration curve. The excitation circuit should have a shunt with a resistance of 10,000 to 20,000 ohms, in order to minimize variations in the resistance of the tissue studied. The galvanic threshold is first determined by establishing the minimum voltage that stimulates when the current lasts an indefinitely long time. Lapicque has given the name "rheobase" (base of current) to this voltage. The minimum time that the rheobase voltage must last to stimulate is called by Lapicque the "utilization time." It is unnecessary to prolong the current longer than this utilization time because the voltage cannot be reduced however long the current

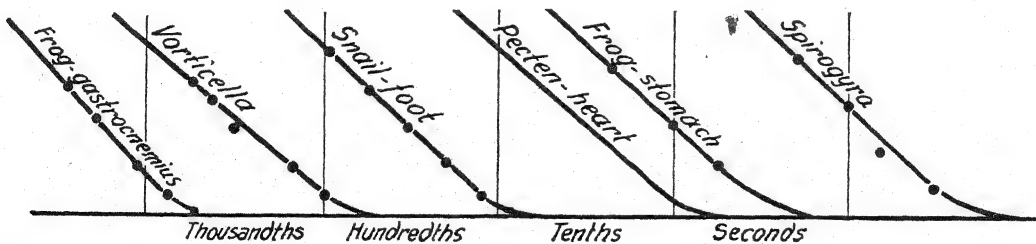



FIG. 324. Chronaxie. Strength-duration curves of animal and plant tissues. All have the same shape, but the time scale varies. (*Lapicque.*)

lasts. The utilization time is of physiologic significance, because a stimulus provokes excitation as soon as it has lasted this time. If the current is shortened to less than the utilization time, it will be necessary to increase its strength in order to obtain a response. There is a utilization time for each voltage, which is longer as the rheobase is approached and shorter as the strength of the stimulus increases. The curve approaches the axis of the strength values asymptotically. In practice, however, if the duration is shortened below a minimum, however high the voltage, there is no stimulation (Figs. 322 and 323).

The utilization time of the rheobase is difficult to determine accurately, because in this section of the curve its variation with respect to the values of strength of stimulus is minimal; therefore a small error in establishing the rheobase may result in a great difference in the time obtained. Lapicque has proposed that the time of a stimulus of twice the strength of the rheobase be determined, as the curve varies more rapidly in this section, and therefore its points can be established with greater accuracy. Lapicque also gives the name of "chronaxie" to this time, which he defines empirically as "the minimum time a current must last to obtain a threshold contraction (the minimal response) with a strength of stimulus (voltage) twice that of the rheobase."

The excitability of a tissue can be determined in practice by measuring its chronaxie. It should not be forgotten, however, that chronaxie is an empirical value; therefore comparable results will be obtained only if both the rheobase and chronaxie are measured in well-defined, standard conditions. Otherwise changes in the conditions of stimulation, rather than changes in excitability, will be measured.

There are therefore two time factors of importance in the development of the excitatory state: (*a*) the duration of the variable phase (rate of change—Du Bois Reymond's law of excitation); (*b*) the duration of the constant phase, or utilization time (Lapicque's chronaxie).

Local and propagated excitatory states. A normal tissue at rest is in condition to be excited; the stimulus changes this condition (excitability) into an excitatory state. A process is started at the point stimulated which increases very rapidly; when it has reached a certain size and covered a certain area, it spreads throughout the whole cell. These two phases are known as (a) the *local excitatory state;* (b) the *propagated excitatory state.* As soon as the local excitatory process begins to be produced by an applied current, a reaction takes place in the excitable substance and another process opposes the excitatory change. This reaction has been called "accommodation" by Nernst. The process of accommodation is slow; in myelinated nerves it has only one-hundredth to one-tenth the speed of the local excitatory state. Accommodation is therefore of little or no importance when stimuli of great strength, and consequently short utilization time, are applied; but it is important when the stimulus has a long utilization time, such as occurs with weak stimuli of a strength near the rheobase and with stimuli consisting not in a sudden but in a slow variation of energy potential.

Latent addition. A subliminal stimulus has no apparent effect, but it must have a hidden one, since a response can be obtained if a sufficient number of these stimuli are sent at adequate intervals. The effect of each successive stimulus is added to that of previous stimuli until the threshold has been attained; there has been "latent addition," as Richet called this process.

In latent addition four factors, varying in relation to each other, must be considered: (*a*) strength of stimulus, or voltage; (*b*) the duration of the constant phase, or chronaxie (it would be better to say "utilization time," since the chronaxie is the utilization time of a stimulus twice the strength or voltage of the rheobase or galvanic threshold); (*c*) the interval between the successive stimuli, or its reciprocal, the frequency of stimulation; (*d*) the number of stimuli needed to attain the threshold. By maintaining constant two of these factors, variations of the other two as functions of each other can be studied.

With constant frequency and number of stimuli, as the duration of each stimulus decreases, the voltage must increase to evoke a response. The strength-duration ratio gives a chronaxie curve. The rheobase is the minimum voltage which provokes a response with a given number of stimuli of long duration discharged at a given frequency. If this voltage is doubled, the minimum duration of each stimulus which evokes a

response when discharged the same number of times at the same frequency is the chronaxie for repetitive stimulation.

With a constant number of stimuli, each one of the same duration, as the interval between the stimuli increases (or the frequency diminishes) the voltage must be increased to obtain a response. If the interval between stimuli is lengthened excessively, the voltage will be that of the threshold for a single stimulus, because the effect of each subliminal stimulus will have completely disappeared before the following stimulus is applied. If, on the other hand, the interval is shortened beyond a certain limit, the voltage cannot be reduced, and the minimum remains constant (Fig. 325).

With constant frequency and duration of stimuli, as the voltage is diminished the number of stimuli needed to obtain a response increases. The curve that expresses the ratio of these two factors is an asymptote approaching the axis of voltage with small numbers of stimuli, and the axis of number of stimuli with low voltages. With a given voltage and duration, the product of the interval between the stimuli (or its reciprocal, *i.e.*, frequency of stimuli) by the number of intervals (number of stimuli minus 1) is a constant, which Lapicque calls "summation time." Repetitive excitability can be defined by two time factors, chronaxie and summation time; these increase and diminish simultaneously.

These observations show that the local excitatory state takes some time to develop and attain the threshold that fires off a propagated excitation. The size of this local excitation is dependent on the strength of stimulus; it starts to disperse as soon as it commences and finally disappears completely. Summation takes place when there remains a sufficient amount of the local excitatory state produced by a subliminal stimulus at the time the following stimulus is applied.

Repetitive excitability. A large number of tissues and functional systems are only slightly or not at all sensitive to a single stimulus. However great the strength and duration of a single stimulus, it may evoke no response or only a small one in this type of effector, *e.g.*, glands and smooth muscle. To obtain the full response, stimuli must be applied repeatedly with an adequate frequency and for a sufficient time. Lapicque has called this type of response "repetitive excitability." It is of great interest because the physiologic excitation of neurons takes place by repetitive stimulation.

Wedensky inhibition. When a nerve is continuously stimulated at relatively high frequency, after a time the response diminishes and eventually ceases, *e.g.*, tension developed by a muscle in tetanus (sustained contraction) subsides gradually until the muscle is completely relaxed. This is known as Wedensky inhibition. At one time it was considered as due to incomplete recovery between the successive stimuli and prolongation of the refractory period (see below), because a response can be obtained by stimulating at lower frequency. There are, however, several phenomena included under the single term of "Wedensky inhibition." One type of depression occurs after relatively prolonged stimulation; it is due to the release of an insufficient amount of chemical mediator, acetylcholine, in muscle. In this case the injection of a small amount of acetylcholine, not large enough to evoke a response, will summate its effect with that of the stimulation which has become subliminal, and the effector will again respond. In other cases depression is due to an excess of chemical mediator accumulated during excitation at high frequency, or because the mediator is not destroyed. This effect can be produced by injecting large doses of the mediator, or by

FIG. 325. Voltage and frequency of stimuli needed to obtain summation. Forty-eight stimuli of $\frac{1}{60}$ and $\frac{1}{90}$ of a chronaxie were applied to the columella muscle of the snail *Helix*. (*Chauchard, A., and B. Chauchard, Ann. de. Physiol., vol. 1, p. 6, 1925.*)

protecting it from destruction[1] (see "Chemical phenomena of neuromuscular transmission," Chap. 67).

The refractory period. As soon as the propagated excitatory state is produced, the tissue becomes inexcitable, and for a short time no response can be obtained even when a stimulus of great strength is applied. This is known as the *absolute refractory period.* It is followed by a second phase, called the *relative refractory period*, during which excitability is gradually restored. The strength of the stimulus needed to evoke a response is at first considerable; then it diminishes progressively until the normal threshold is recovered. In certain conditions a period of hyperexcitability follows the relative refractory period (Fig. 326). The capacity of response during this relative refractory period is depressed; *e.g.*, the nerve fiber conducts at lower speed an impulse that has a smaller action potential.

FIG. 326. Recovery of excitability after excitation in a nerve-muscle preparation of a frog. (*Adrian and Lucas, J. Physiol., vol. 44, p. 68, 1912.*)

The duration of the refractory period varies in different tissues and depends on the state of excitability. Bowditch was the first to observe the refractory period when studying the excitability of heart muscle. The heart has a very long refractory period, which lasts throughout the phase of contraction. In nerve fibers of the highest speed of conduction, the absolute refractory period lasts from 0.4 to 0.5 msec., and the threshold returns to normal in 2 to 3 msec. Striated muscle in the cat, stimulated through the nerve, has a refractory period of 3 msec.; when the muscle is stimulated directly this is reduced to 2.2 msec.; the muscle end-plate, therefore, prolongs the refractory period by 0.8 msec.

[1] ROSENBLUETH, A., K. LISSÁK, and A. LANARI, *Am. J. Physiol.*, **128**, 31, 1939.

Rosenblueth and his associates[1] maintain that the division of the refractory period into absolute and relative has no physiological significance; there is a *functional refractory period*, *i.e.*, the time after a conditioning stimulus during which the action potential developed by a second stimulus at the site of excitation is not of sufficient amplitude (voltage) to reach the threshold of the neighboring parts of the nerve.

The all-or-nothing law. Bowditch (1871), working in Ludwig's laboratory, showed that a frog's heart, brought to a standstill by a Stannius ligature, responded to a threshold stimulus with a maximal contraction; no greater response could be obtained by increasing the strength of the stimulus, other conditions being equal. "The reason why the heart can vary its degree of contraction must be sought in the fibers themselves and not in the changes of the stimulating agent."[2] Gotch obtained evidence showing that the changes in response of striated muscle, stimulated directly or through the nerve, with stimuli of different strengths, were due to the activation of a variable number of contractile units, each one of which gave a maximal contraction. Pratt and Eisenberger[3] gave direct and definite proof that each muscle fiber responds maximally to the threshold stimulus. They used a microelectrode and registered the response of individual fibers. They observed that when a fiber contracts it does so to the utmost of its capacity under the existing conditions (Fig. 327).

"All or nothing seems to be the motto of the heart when stimulated," was Ranvier's commentary on Bowditch's work, and this dictum includes striated muscle and nerve. The lack of relationship between size of response and strength of stimulus is easily seen in the heart because it contracts as a single unit. Striated muscle is made up of many units, which can respond separately; contraction increases with the strength of the stimulus because a larger number of units is thus excited. It must be borne in mind, however, that the response of a given unit can vary when the conditions of the tissue are changed.

Stimulation of an isolated fiber, with an electrode sufficiently large to act upon the whole

[1] ROSENBLUETH, A., *et al.*, *J. Cell. & Comp. Physiol.*, **33**, 405, 1949.

[2] BOWDITCH, H. P., *Ber. Sächs. Ges.*, **25**, 652, 1871.

[3] PRATT, F. H., and V. P. EISENBERGER, *Am. J. Physiol.*, **49**, 1, 1919.

fiber, has shown that it is possible to obtain graded contractions by increasing the strength of the stimulus. Tension developed is a function of voltage when stimuli of constant duration are used and a function of duration when the voltage remains constant. Stimulation with a very fine electrode provokes the contraction of only that part of the fiber which is near the electrode. A stronger stimulus increases the contracting area. A still stronger stimulus will produce a contraction that spreads to the whole fiber and is of an all-or-nothing character, *i.e.*, the weakest stimulus capable of producing this propagated disturbance evokes a maximal contraction. Both the local and the propagated excitation give normal strength-duration curves (chronaxie curves), but only the propagated excitation is accompanied by a "spike" potential.[1] As long as the muscle fiber remains in good condition, it is very difficult to obtain a localized response graduated in relation to the strength of the stimulus. This is made easier by the use of anesthetics, an excess of K^+, or any other agent that damages the integrity of the fiber. In physiologic conditions striped muscle, nerve, and heart give the maximal response with the minimum stimulus that can provoke a propagated excitation.

[1] The local excitatory state is also accompanied by electrical phenomena, but there is not a large "spike" potential such as is seen in the propagated excitatory state (see "Electrical phenomena of muscular activity," Chap. 67).

Adrian has pointed out that an all-or-nothing response is a consequence of a threshold and a refractory period. When the local excitatory process reaches the threshold, a propagated disturbance is initiated and excitability disappears for a short time (absolute refractory period). Therefore, even if the stimulus continues it cannot produce a greater effect because the tissue is incapable of response (Fig. 328).

FIG. 327. The "all-or-nothing law." The lower tracing registers the strength of the stimulus; the upper tracing, the response of the muscle. The height of contraction increases suddenly, *e.g.*, at *A* and *B*, when the gradual increase in strength of stimulus reaches certain levels. (*Pratt, F. H., and V. P. Eisenberger, Am. J. Physiol., vol. 49, p. 1, 1919.*)

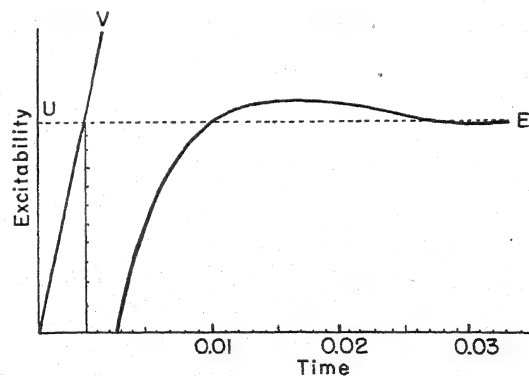



FIG. 328. Changes in excitability produced by stimulation. When the local excitatory state *V* produced at the cathode reaches the threshold *U*, the propagated excitatory state is fired off and excitability *E* disappears completely (absolute refractory period); later excitability is gradually recovered during the relative refractory period. Persistence of *V* above the threshold cannot evoke a second impulse because excitability has been abolished.

The local excitatory state is not all-or-nothing in character. Its size depends on the strength of the stimulus, or better still on the relation between the strength of the stimulus and the excitability of the tissue. It is not followed by a refractory period and does not have a spike potential. As soon as this state attains a certain threshold size it starts a propagated disturbance, which is all-or-nothing in character, is followed by a refractory state, and has a spike potential.

Excitability of different tissues. Though all cells and tissues are excitable and there are laws common to the excitability of all of them, there are also differences between the tissues. Thus, generally speaking, tissues that respond slowly have a higher rheobase and a longer chronaxie than tissues that respond rapidly. The frog gastrocnemius has a chronaxie of 0.3 msec. and a contraction time of 0.15 sec.; gastric muscle in the same species has a contraction time of from 15 to 20 sec. and a chronaxie that varies between 30 and 100 msec.; the pigment cells (chromatophores) in the frog's skin react very slowly, and their chronaxie is 11 to 15 sec. (Table 85).

Factors that modify excitability. When studying excitability it should be kept in mind that currents of short duration excite only at the cathode; therefore only changes taking place under the cathodic electrode produce an effect. Generally speaking, rheobase and chronaxie vary in inverse direction.

Cold diminishes the threshold (rheobase) and increases chronaxie; if sufficiently intense it will abolish excitability. A change of the *acid-base equilibrium* toward the alkaline side diminishes the rheobase and increases chronaxie; one toward the acid side produces opposite effects. Alkalosis produces an increase in muscle tone and spontaneous contractions that can become convulsions; this syndrome is called tetany. In acidosis, muscles are relaxed and excitability is depressed. Excitability is also conditioned by the *ion equilibrium.* Monovalent and bivalent cations are antagonistic to each other. Variations in the ratio of $\frac{K^+ + Na^+}{Ca^{++} + Mg^{++}}$ cause changes in excitability. A relative or an absolute increase in monovalent cations lowers the rheobase (threshold) and increases the chronaxie; the predominance of bivalent cations has the opposite effect. Removal of the parathyroid glands diminishes the blood calcium (hypocalcemia); therefore the threshold of excitability is lowered and tetany occurs. Hypercalcemia, on the contrary, depresses excitability. Application of a constant electric current provokes changes in excitability which will be discussed later (see "Electrotonus").

Table 85. Excitability of Different Tissues

Tissue	*Chronaxie, msec.*	*Utilization time, msec.*	*Contraction time, sec.*
Frog gastrocnemius...	0.3	3	0.1
Tortoise heart.......	8.2	80	2–3
Frog stomach........	30–100	...	15–20
Adductor muscle of crab's claw........	30	300	5
Frog chromatophores..	11,000–15,000		

ELECTRICAL PHENOMENA IN TISSUES

Toward the end of the eighteenth century Galvani demonstrated the existence of electrical phenomena in living tissues. He exposed the sciatic nerve of a frog and placed it on the leg muscle of another frog, having first injured this muscle. Each time the nerve closed a circuit between the undamaged surface and the injured part of the muscle, the muscles innervated by the nerve contracted. There is a difference in

FIG. 329. Galvanoscopic leg. Each time the nerve *nb* of the muscle *B* closes the circuit between the normal surface 1 and the injured surface 2, *B* contracts.

potential between the intact and the injured parts of the muscle; the nerve acts as a conductor and the electric current thus set up stimulates it. The 'galvanoscopic leg" was the first galvanometer to be invented. It was the means of registering for the first time the effects of a direct electric current (Fig. 329).

Demarcation potential and action potential. When a nerve or muscle is injured and the damaged part is joined to the intact surface by a circuit in which there is a sufficiently sensitive galvanometer, the injured part is found to be electronegative with respect to the parts that are undamaged (Matteuci, 1838). This is the injury or demarcation potential; it diminishes in approximately exponential form with distance from the injury (Fig. 330 I). It is due to the existence of an electromotive force in the intact membrane, the *membrane resting potential*, the protoplasm being negative and the external fluid positive. Injury removes the electromotive force, and there is no potential difference at the junction of protoplasm and external fluid. The demarcation potential sets up a flow of current, the demarcation current, through a considerable segment of the fiber (Fig. 330 II).

When a cell is stimulated and an excitatory state is evoked, the point stimulated becomes electronegative with respect to the parts at rest (Du Bois Reymond). A "negative variation" spreads like a wave over the whole surface of the cell (Bernstein); this is the *action potential*, or more precisely the spike potential, of the propagated excitatory state. It travels at a definite

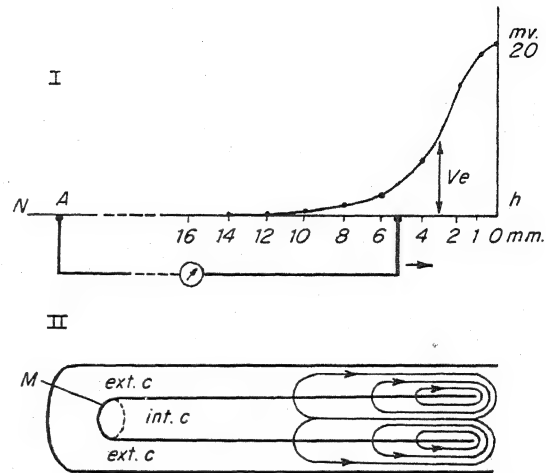



Fig. 330. I, longitudinal distribution of the demarcation potential *Ve* along the nerve *N*: *A*, a point of the nerve at which there is no demarcation potential; *h*, point of injury. II, demarcation current: *M*, membrane; *ext. c.*, external conductor; *int. c.*, internal conductor (core). (*Lorente de Nó, R., Studies of the Rockefeller Inst. Med. Res., vol. 131, New York, 1947.*)

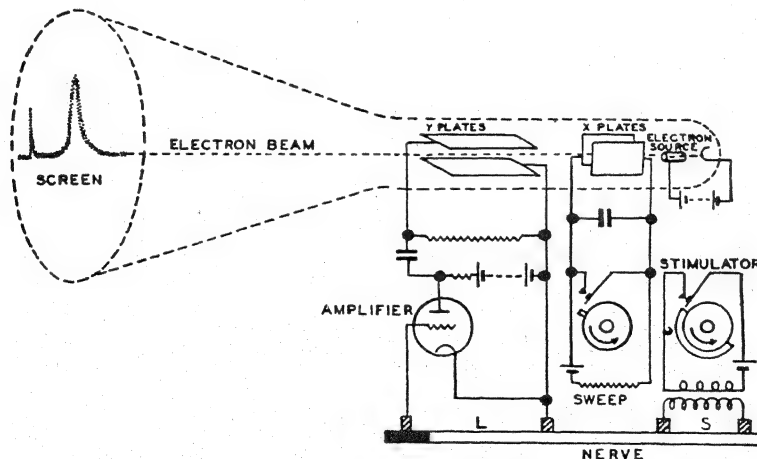



Fig. 331. Simplified diagram of cathode-ray oscillograph. *S*, stimulating electrodes, through which an induction shock, or a condensor discharge, can stimulate the nerve; *L*, electrodes of registering circuit, one terminal on the intact surface, the other on the killed end of the nerve; thus a monophasic variation is obtained. The source of electrons (cathode rays) is a Braun tube; the rays are "focused" on the fluoroscope or the photographic paper of a photokymograph. As the electrons pass between the *X* plates, the rays can be moved in a horizontal plane by changes in potential in the sweep circuit; as they pass between the *Y* plates, they can be similarly moved in a vertical plane. When the nerve is stimulated at *S* the action potential on reaching the circuit *L* provokes an oscillation of the cathode rays as they pass through *Y*. (*Erlanger, J., and H. S. Gasser, "Electrical Signs of Nervous Activity," University of Pennsylvania Press, Philadelphia, 1937.*)

speed—100 m/sec. in the fastest mammalian nerve fibers, and only a few centimeters per second in the slowest fibers.

Resting and action potentials are of very low voltage; the convenient unit of measure is the millivolt (1 mv. = 0.001 volt). Vacuum-tube amplifiers have been used to amplify potential variations of any form without distortion, thus overcoming all difficulties of registration due to the low voltage. Action currents are also of very short duration; the unit of time convenient for their measurement is the millisecond (1 msec. = 0.001 sec.). The inertia of the recording apparatus is a serious obstacle for the accurate study of these phenomena. One of the first instruments used was Lippmann's capillary electrometer. Later, Einthoven's string galvanometer became the standard instrument for the study of electrical variations in the heart (see Chap. 15, The Electrical Phenomena of Cardiac Activity). Braun's cathode-ray tube has been adapted by Erlanger and Gasser for use as an oscillograph and can register changes in potential that take place in 0.1 msec.; the results obtained are accurate to 0.02 msec. (Fig. 331).

FIG. 332. Methods for obtaining absolute values of resting and action potentials in single fibers. *A*, external electrodes applied to single fiber; *i*, crushed nerve; *s*, thin layer of saline; *o*, oil. *B*, longitudinal insertion of electrode *e*, filled with sea water, into giant axon of squid; *w*, sea water. *C*, transverse insertion of internal glass capillary electrode *e*, filled with isotonic or 3*M* KCl solution; *R*, Ringer's fluid. (*Hodgkin, A. L., Biol. Rev., vol. 26, p. 339, 1951.*)

Isolated surviving nerve or muscle fibers are often used in the study of electrical phenomena of living tissues. These fibers are either surrounded by air (or another dielectric medium) or submerged in a solution of dissociated electrolytes. The former, although an artificial procedure, is a most useful one, but in the living organism excitable tissues are in an environment which is essentially a solution of dissociated electrolytes and therefore conducts electricity. A difference in potential between two points in a medium of this nature causes a flow of current from the positive to the negative pole throughout the whole medium, if it is not too extensive. If a nerve or muscle fiber surrounded by a dielectric medium is stimulated, no change is recorded in a part at rest until the disturbance reaches it. If the nerve is surrounded by a conducting medium there is a fluctuating field of current flow throughout the medium while the disturbance is being propagated; a part of the nerve that is at rest acts as a source of current flow until the propagated disturbance reaches it, as a sink of current flow while activity lasts, and again as a source when the disturbance spreads beyond it.

The resting and action potentials of single fibers can be recorded by several methods:

1. Electrodes are placed on the surface of the fiber or cell (external electrodes); the potentials recorded are reduced by short-circuiting in the external fluid, but this reduction is corrected by determining the ratio between the internal resistance of the fibers and the resistance of the external fluid[1] (Fig 332*A*).
2. A capillary electrode filled with sea water, or isotonic KCl, is introduced into one end of a fiber and pushed into it along its axis for 10 to 30 mm.; the second electrode is placed in the external fluid. The effects of damage made by introduction of the electrode do not extend beyond 10 mm. from the point of insertion. This method has been used in giant plant cells and axons of squids with fiber diameters of 200 to 500 μ[2] (Fig. 332*B*).
3. Electrodes with a tip diameter less than 0.5 μ can be introduced transversely through the membrane without damaging it; they are usually filled with 3*M* KCl in order to reduce junction potentials to a low value. The second electrode is placed in the surrounding fluid[3] (Fig. 332*C*).

[1] KATZ, B., *Proc. Roy. Soc., London, s.B.*, **135**, 506, 1948.
[2] OSTERHOUT, W. J. V., *Biol. Rev.*, **6**, 369, 1931; HODGKIN, A. L., and A. F. HUXLEY, *Nature, London*, **144**, 710, 1939; CURTIS, H. J., and K. S. COLE, *J. Cell. & Comp. Physiol.*, **15**, 147, 1940.
[3] LING, G., and R. W. GERARD, *J. Cell. & Comp. Physiol.*, **34**, 382, 397, 413, 1949; NASTUK, W. L., and A. L. HODGKIN, *J. Cell. & Comp. Physiol.*, **35**, 39, 1950.

The resting membrane potential has been recorded in many excitable tissues. It ranges be-
een 50 and 100 mv. Excitation reverses this
tential by 30 to 50 mv., so that the action
tential ranges from 80 to 130 mv.[1] (Table 86,
;s. 333, 334) (see "Electrical phenomena of

FIG. 333. Diagram of axon membrane potential at rest.

ıscular activity," Chap. 67, and "Electrical
enomena of the nerve impulse," Chap. 68).

Properties of the membrane. Muscle and
rve fibers can be considered as *core conductors*
:ermann), *i.e.*, tubes with a conducting core
›arated from outside conductors by a resistant

ble 86. Magnitude of Resting and Action Potentials

ıimal	*Tissue*	*Fiber diameter, μ*	*Resting potential, mv*	*Action potential, mv*	*Electrodes*
igo..	Unmyelinated nerve	500	50	90	Internal
ia...	Unmyelinated nerve	200	62	120	Internal
ıa...	Myelinated nerve	15	71	116	External
ıa...	Striated muscle	80	88	119	Internal
ıa...	Cardiac muscle	30	50–90	65–115	Internal
ȝ....	Purkinje fiber	30	90	121	Internal
·*...	Spinal motoneuron	...	70	95	Internal

ɔource: HUXLEY, A. F., and R. STÄMPFLI, *J. Physiol.*, , 476, 1951.
*BROCK, L. G., J. S. COOMBS, and J. C. ECCLES, *Physiol.*, **117**, 438, 1952.

d capacitative membrane. The membrane can
represented by an infinite series of elements
parallel connected by intracellular and extra-
lular fluid conductors (Fig. 335). Each ele-
:nt has a condenser (c), a resistance (r), and a
ttery (*emf*) with the positive plate toward the
ter surface. The membrane acts as a rectifier,
and the current necessary to maintain a depolarization is much greater than that needed to produce an equivalent increase in membrane potential.[1] In the resting membrane there is no flow of current because all the elements are at the same potential (Fig. 335).

FIG. 334. Diagram of axon membrane potential conducting an impulse.

The effects of ions on nerve trunks, and the electrical behavior of the latter in response to applied currents, in many respects are not consistent with the core-conductor theory. The resistance of the epineurium is the cause of these phenomena.[2] If the epineural connective sheath is removed, the nerve follows closely the expectations of the theory, and restoration of the sheath reproduces the original permeability and electrical properties of the nerve.

The membrane has selective permeability, *i.e.*, certain substances pass through it more or less easily, and to others it is impermeable. The size of the particle with respect to the size of the membrane pores, and in the case of ions the electric charge, are fundamental factors in conditioning the passage through the membrane. A membrane with selective permeability which separates two electrolyte solutions (protoplasmic and interstitial fluid) is polarized; ions are distributed on its internal and external surface in an "electric double layer" (Helmholtz). In excitable tissues cations form a positively charged layer on the outside surface and anions a negatively charged layer on the inside surface (Fig. 333).

Lorente de Nó[3] postulates three fractions in the membrane potential, Q, M, and L, maintained at three concentric double layers, q-m, m-m', and m'-l; q is the most external, and l is supposed to be situated in the core. Apparently there is no potential difference

HODGKIN, A. L., *Biol. Rev.*, **26**, 339, 1951.

[1] COLE, K. S., and J. S. CURTIS, *J. Gen. Physiol.*, **24**, 551, 1945.

[2] RASHBASS, C., and W. A. H. RUSHTON, *J. Physiol.*, **109**, 327 and 343, 1949; **110**, 110, 1949; RUSHTON, W. A. H., *J. Physiol.*, **109**, 314, 1949; RASHBASS, C., *J. Physiol.*, **109**, 354, 1949.

[3] LORENTE DE NÓ, R., "A Study of Nerve Physiology," Studies of the Rockefeller Inst. Med. Res., 131 and 132, New York, 1947.

at the boundary between *q* and the external fluid, because substitution of the ions in the external fluid by sugars, or of the cations by choline, does not produce significant changes in the membrane potential. The *Q* fraction is large and rapidly restored (*Q* = quick) during the descending phase of the spike potential. The *M* and *L* fractions are smaller and are restored more slowly during the negative after potential. The *L* fraction is very sensitive to external influences (*L* = labile), such as cooling and ether anesthesia; it is small in a nerve maintained in oxygen but increases rapidly and considerably if O_2 is replaced by 5 per cent CO_2 and 95 per cent O_2.

Fig. 335. Diagram of electrical structure of excitable membrane.

Electrotonus. When a galvanic current is applied to an excitable tissue (polarizing current), part of it flows from one electrode to the other along the conductors outside the membrane, and part through the membrane and the cytoplasm, not only between the electrodes (intrapolar segment) but also spreading in pseudowavelike form in the extrapolar regions laterally to the cathode and anode (Fig. 336). Excitable tissue does not behave as a passive conductor but reacts to the applied current, and electromotive changes take place which have been called "electrotonus" by Du Bois Reymond. The potential (electrotonic potential) established by the flow of current across the membrane is a variation of the membrane potential. This potential spreads in pseudowavelike form along the extrapolar segments, diminishing exponentially as the distance from the electrodes increases.

Fig. 336. Polarization of nerve *N* with galvanic current. P_1P_2, polarizing circuit; R_1R_2, recording circuit.

Lorente de Nó[1] distinguishes several components in the electrotonic potential, E_1, E_2, E_3, E_4. Each one has characteristic electrical features, and each to a certain extent may vary independently of the others. These components are found in the cathode and anode and are related to the different fractions of the membrane potential. The E_1 component develops rapidly and does not spread far; it is a variation of the *Q* fraction of the membrane potential ("fast" electrotonus). E_2 and E_3 are variations of the *M* and *L* fractions respectively; they develop less rapidly ("slow" electrotonus) but spread more, so that at 15 mm. from the polarizing electrode the electrotonic potential is mostly "slow" electrotonus. The E_4 deflection is related to postcathodal depression (see below).

Electrotonus and excitability. Application of a constant current tends to decrease membrane polarization at the cathode and to increase it at the anode while the current is flowing; the reverse occurs when the current ceases to flow.

[1] *Loc. cit.*

This causes considerable changes in excitability during and after the flow of current. Changes in excitability are, in general, parallel to variations in the electrotonic potential. A negative variation lowers and a positive variation raises the threshold.[1] At the cathode (cathelectrotonus) during the flow of current excitability rapidly rises to a maximum, then falls gradually while the current is still flowing. After the current ceases to flow excitability is depressed below the resting level (postcathodal depression). At the anode (anelectrotonus) excitability rapidly falls below the resting level and then rises gradually. When the flow of current ceases, excitability rises above the resting level (Fig. 337).

Cathodal and anodal block. Propagation of the excitatory state (*e.g.*, conduction of a nerve impulse) can be blocked by the polarizing current both at the cathode and at the anode. Cathodal block is due to depolarization and is reinforced by repetitive stimulation. Thus, a single impulse, or a series of impulses at long intervals, may pass through a cathodal region which blocks a series of impulses of higher frequency. Anodal block is caused by an increase in polarization, and impulses arriving at the anodal region do not reinforce, but on the contrary diminish, anodal block; thus a single impulse may be stopped, but a series of impulses may pass through if delivered at a sufficiently high frequency to permit summation of depolarization produced by each impulse.

[1] Fluctuations of E_1 and E_3 produce large changes, and those of E_2 small changes, in excitability.

FIG. 337. Changes in excitability (thin line) and electrotonic potentials, *P*, produced at 6 mm. from the polarizing electrode. On the left, cathodal current; on the right, anodal current. Above, beginning of flow of polarizing current; below, end of flow. (*Lorente de Nó, R., Studies of the Rockefeller Inst. Med. Res., vol. 131, New York, 1947.*)

THE MECHANISM OF MEMBRANE POTENTIALS

The role of metabolism. Membrane potentials of excitable cells can be kept up for a time in tissues separated from the organism and in the absence of oxygen. Anoxia, however, gradually produces depolarization and the membrane potential declines. If oxygen is again available repolarization rapidly takes place and the membrane potential is restored. Even prolonged anoxia does not, however, cause complete depolarization of nerves. A residual potential remains which declines at a very slow rate, and depends on oxidative metabolism only in so far as metabolism is necessary for the maintenance of the normal structure of the fiber. This potential is probably due to differences in the concentration of ions inside and outside the fiber.

The manner in which oxidative metabolism maintains the resting potential is not known. It was usually supposed that metabolic processes acted indirectly, being necessary for the maintenance of the normal structure of the excitable substance and membrane permeability. A more direct action has also been

postulated. According to Lorente de Nó[1] the membrane potential is the result of an equilibrium between electrostatic forces of attraction in the double layer (which would collapse if these forces alone existed) and electromotive forces in the membrane which tend to separate the charged particles of opposite sign. These electromotive forces are established by electrochemical reactions dependent on oxidative metabolism. It has also been postulated that metabolic processes are necessary for the active extrusion of Na from the cells, which may leak in during rest and which enters in large quantities during activity (see below).

The role of ions. Bernstein,[2] on the basis of the fact that potassium concentration is much greater in the excitable cells than in the external fluid, and supposing the membrane to be impermeable to Na^+ and anions, attributed the membrane potential to the difference between the concentration of K^+ inside and outside the cells. Further observations have shown that this theory is inadequate, but undoubtedly the resting and action potentials depend on (*a*) the concentration of ions within and outside the cell; (*b*) the permeability of the cell membrane; (*c*) the metabolic processes which condition ionic concentrations and membrane permeability, discussed in the preceding paragraph.

The concentration of ions in cells and tissue fluids and movements of ions through cell membranes can be accurately determined. The development of sensitive micromethods of analysis and the use of radioactive isotopes have been of great value in this respect. These methods, however, do not give much information on the speed of chemical changes and therefore on the phase of a rapid physiologic phenomenon (*e.g.*, excitation) in which they occur.

Ionic concentrations and membrane permeability in the resting state. The concentration of *potassium* in excitable cells is 20 to 50 times the concentration in the external media (Ringer's fluid, sea water, plasma, or dyalized blood).[3] Most of the potassium (80 to 90 per cent) in nerve and muscle fibers is free to exchange and probably exists as free ions.[4] *Sodium* is three to fifteen times more concentrated in the external fluid than in the cells, and *chloride* five to fifty times more concentrated outside than within the cells. In excitable tissues, therefore, there is high K^+ and low Na^+ and Cl^- concentration in the cells. The deficit of inorganic anions is considerable; in muscle it is more than 120 mEq./kg. of fiber water.[1] Electrical neutrality is maintained by organic anions. Proteins and phosphate compounds are found in high concentration in vertebrate muscle, but the nature of organic anions which balance K^+ has not been established.

According to Conway[2] and his associates the size of the hydrated ions determines whether or not they can pass through the membrane. The membrane is relatively permeable to K^+ and Cl^-, and their equilibrium concentrations are in agreement with the Donnan relation: $[K]_i[Cl]_i = [K]_0[Cl]_0$. Thermal agitation and potential differences are the forces that move these ions through the membrane. Experiments with radioactive isotopes (K^{42} and Cl^{38}) have given results that, with certain restrictions, are in agreement with the Conway hypothesis. The hydrated sodium ion is larger than the hydrated potassium ion, and the resting cell membrane is relatively impermeable to it. A study of the movements of Na^+, made with a radioactive isotope (Na^{24}), has shown[3] a ratio of inward to outward flux of Na^+ that has led to the hypothesis of the existence of an active mechanism which expels from the cell sodium that leaks into it. This "sodium pump" is dependent on a metabolic process.

The magnitude of the resting membrane potential calculated from the potassium concentrations inside and outside the fiber, taking into account Cl^- and other ions, is in good agreement with the figures obtained by direct measurement. If potassium concentration increases in the external fluid, it has a depolarizing action and the membrane potential decreases proportionally to the K^+ concentration. The fall in membrane potential is such as may be expected if the membrane is permeable to K^+ and Cl^- and very little to Na^+ and other ions. There is therefore a considerable amount of evidence for the

[1] *Loc. cit.*

[2] BERNSTEIN, J., "Elektrobiologie," Friedrich Vieweg & Sohn, Brunswick, 1912.

[3] FENN, W. O., *Physiol. Rev.*, **16**, 450, 1936; BOYLE, P. J., and E. J. CONWAY, *J. Physiol.* **100**, 1, 1941.

[4] KEYNES, R. D., and P. R. LEWIS, *J. Physiol.*, **113**, 73, 1951; KEYNES, R. D., *J. Physiol.*, **113**, 99, 1951.

[1] HODGKIN, A. L., *Biol. Rev.*, *loc. cit.*

[2] BOYLE and CONWAY, *loc. cit.;* CONWAY, E. J., *Nature. London*, **157**, 715, 1946.

[3] KEYNES, *op. cit.*, **114**, 119, 1951.

importance of K^+ concentration for the establishment of the membrane potential.[1]

The action potential. When an excitable cell is stimulated and enters into activity, the membrane potential is reversed. The inside of the resting cell is negative to the external fluid by 50 to 100 mv.; at the height of activity it is 30 to 50 mv. positive (Table 86 and Figs. 333 and 334). This reversal is attributed to a sudden change in the membrane which becomes highly and specifically permeable to sodium during the rising phase of the spike potential. During the falling phase permeability for potassium increases and the cell loses K^+. A great deal of experimental evidence has been brought forward in support of this hypothesis:

1. The inflow and outflow of sodium have been measured in giant axons of squids by activation analysis and by means of a radioactive sodium isotope (Na^{24}).[2] There is a large increase in the inflow and outflow of sodium, with a net rise in intracellular Na^+ sufficient to account for the transfer of charge which must take place to reverse the potential.
2. It has long been known that muscles immersed in sodium-free Ringer's fluid lose their excitability (Overton). More recent work has confirmed and extended this observation, showing that nerve fibers do not conduct when submerged in a fluid free from sodium, *i.e.*, the excitatory state is not propagated along such a fiber.

[1] Lorente de Nó (*loc. cit.*) does not agree with this conclusion. He finds that anoxic depolarization and recovery in oxygen are not changed significantly by the absence of K^+ in the external fluid.

[2] KEYNES, *op. cit.*, **114**, 119, 1951; ROTHENBERG, M. A., *Biochim. et biophys. acta*, **4**, 96, 1950; GRUNDFEST, H., and D. NACHMANSOHN, *Federation Proc.*, **9**, 53, 1950.

3. If the sodium concentration in the external fluid is reduced, the resting potential increases slightly and the action potential diminishes considerably. The potential is no longer reversed on excitation when the external sodium concentration is one-third to one-sixth the normal. Moreover the maximum rate of rise of action potential is proportional to the external sodium concentration.[1]
4. Potassium inflow and outflow are increased by stimulation, and nerve and muscle fibers lose K^+. A nerve fiber loses at each impulse an amount of K^+ equivalent to that of Na^+ entering it.[2]
5. There is a considerable decrease in the resistance of the membrane during activity,[3] which falls from a resting value (in giant axons of squids) of about 1,000 ohms/sq.cm. to 25 ohms/sq. cm. during activity. Membrane conductance increases rapidly during the rising phase of the spike potential and reaches a high value in less than 0.1 msec. It then decreases but remains above the resting level during the positive phase of the action potential and the early part of the refractory period (Fig. 338). The capacity of the membrane does not change significantly during activity; it may fall 1 to 2 per cent from the resting value of 1.5 μf/sq. cm. (squid axon).

FIG. 338. Diagram of electrical changes during excitation. Conductance (solid line) and electrical potential (broken line) in a giant axon of a squid on the passage of an impulse. (*Cole and Curtis, J. Gen. Physiol., vol. 22, p. 649, 1939.*)

[1] HODGKIN, A. L., and B. KATZ, *J. Physiol.*, **108**, 37, 1949; NASTUK, W. L., and A. L. HODGKIN, *J. Cell. & Comp. Physiol.*, **35**, 39, 1950; HUXLEY, A. F., and R. STÄMPFLI, *J. Physiol.*, **112**, 496, 1951.

[2] HODGKIN, A. L., and A. F. HUXLEY, *J. Physiol.*, **106**, 341, 1947; KEYNES, *op. cit.*, **113**, 99, 1951; **114**, 119, 1951.

[3] CURTIS, H. J., and K. S. COLE, *J. Gen. Physiol.*, **21**, 757, 1938; COLE, K. S., and H. G. CURTIS, *J. Gen. Physiol.*, **22**, 649, 1939; KATZ, B., *J. Neurophysiol.*, **1**, 169, 1942; TASAKI, J., and K. MIZUGUCHI, *Biochim. et biophys. acta*, **3**, 484, 1949.

The process of excitation. When a stimulus, *e.g.*, an electric shock, is applied to an excitable tissue, the following events occur: The permeability of the membrane changes abruptly so that sodium ions enter into the cell, the membrane is depolarized, and an electrotonic potential is set up. If this local disturbance occupies a sufficiently large area, the threshold of a propagated disturbance is reached, local currents flow into the active region from the neighboring areas, which are thus depolarized, and excitation spreads along the surface of the cell, accompanied by a spike potential. Na^+ inflow takes place during the first part of the process and the membrane potential is reversed; K^+ flows out during the declining phase of the spike potential. During the recovery period the membrane is repolarized, but the cell cannot respond to stimulation (refractory period) until the process of recovery (repolarization) has reached a certain level. The role of chemical factors in excitation will be discussed in the following chapters.

BIBLIOGRAPHY

BLAIR, H. A., Conduction and Synaptic Transmission in the Nervous System, *Ann. Rev. Physiol.*, **12**, 399, 1950.

BROOKS, C. McC., and M. G. P. FUORTES, Excitation, Conduction and Synaptic Transmission in the Nervous System, *Ann. Rev. Physiol.*, **14**, 363, 1952.

BULLOCK, T. H., Conduction and Transmission of Nerve Impulses, *Ann. Rev. Physiol.*, **12**, 261, 1950.

DAVIS, H., and A. FORBES, Chronaxie, *Physiol. Rev.*, **16**, 407, 1936.

ECCLES, J. C., Conduction and Synaptic Transmission in the Nervous System, *Ann. Rev. Physiol.*, **10**, 93, 1948.

HODGKIN, A. L., The Ionic Basis of Electric Activity in Nerve and Muscle, *Biol. Rev.*, **20**, 339, 1951.

LAPICQUE, L., "L'excitabilité en fonction du temps," Presses Universitaires de France, Paris, 1926.

LAPICQUE, L., "L'excitabilité itérative," Hermann et Cie, Paris, 1936.

LORENTE de Nó, R., "A Study of Nerve Physiology," Stud. Rockefeller Inst. Med. Res., vols. 131 and 132, New York, 1947.

CHAPTER 67

The Phsiologic Properties of Muscle

MOVEMENT

Movement is one of the most widespread of vital activities; there is no life without it. The protoplasm of the living cell is in constant movement. Its submicroscopic particles show a rapid oscillatory motion known as "brownian movement," there are fluid currents within it, and there is a continual exchange of matter between the cell and its environment. All this activity ceases when the cell dies.

Unicellular animals move from one place to another by means of the so-called "ameboid" movements, or by the activity of flagella or cilia distributed on their outer covering, or under the influence of external agents (tropisms). In animals with a more complex structure (celenterates, arthropods, worms, vertebrates) a special tissue is developed which has the property of contractile movement and which forms the muscles. These organs not only serve for locomotion but also carry out all major displacements of mass within the body.

Ameboid movements. When an adequate stimulus acts on an ameba (or a cell which has ameboid movement) a stream of endoplasm begins to flow toward part of the cell where a salient, called a pseudopod, appears. This is due to contraction of the cortical gel at the rear of the cell. After contracting the gel liquefies and is pushed forward along the central part of the cell toward the pseudopod, where it forms gel on the walls. Several theories have been proposed to explain this type of movement—differences in surface tension, or in pH, or in osmotic pressure between different parts of the cell, rhythmic contraction of the membrane, etc.—but it is not yet clear how these forces act.

Goldacre and Lorch[1] have shown that an injection of 1 to 3 per cent solution of adenosinetriphosphate (ATP) into the tail of an ameba caused contraction to accelerate, and extended wrinkling in the tail region. When ATP was injected into the pseudopod, streaming was reversed. Large amounts of ATP caused rapid putting out and retraction of many pseudopods all over the cell surfaces. These effects lasted less than 1 min. and were again observed on repeating the injection. ATP has an effect on actomyosin similar to its effect on the ameba's tail, *i.e.*, contraction of the actomyosin fiber and liquefaction of the gel (see "Physico-chemical structure of the myofibril"). Goldacre and Lorch suggest that in the tail an "organizer" produces ATP which causes folding of the polypeptide chains of the cortical gel, which thus contracts and liquefies; in the region of the pseudopod the protein unfolds and cortical gel is formed.

The leukocytes of vertebrates progress by ameboid movement at a speed of 30 to 35 μ per min., but for this they must be on a resistant surface; in a fluid medium, such as blood plasma, they have a spherical form. Temperature, saline concentration, ion and acid-base equilibriums, oxygen and CO_2 tensions, several chemical substances, and other factors in the environment condition the ameboid activity of free cells.

Ciliary movement. Cilia are permanent, threadlike protoplasmic organs inserted in the free surface of the cells of certain epithelia and over the whole surface of some free cells (ciliated bacteria and protozoa). Ciliary movement takes place in two stages: (*a*) all the cilia move together in the same direction in a sudden fast movement, like the lash of a whip; (*b*) a slow movement follows which brings the cilia back to the resting position. During the first movement the cilia become rigid, and during the second they are flexible and bend. Particles on

[1] GOLDACRE, R. J., and I. J. LORCH, *Nature, London,* **166**, 497, 1950.

the epithelial surface are displaced in the direction of the rapid movement. Ciliated protozoa progress in a direction opposite to the rapid movement, just as swimmers do. The rate of movement is increased by heat up to an optimal temperature, by alkalinization of the surrounding fluid, and by an increase in K^+ up to a certain concentration. It is decreased, or even stopped, by cold, acidification, Ca^{++}, asphyxia, and anesthesia. The movements of the cilia imply the activity of the whole cell, as is shown by its oxygen consumption, which increases and decreases with the rate of ciliary movement. When the cilia are stimulated at one point, a wave of movement is transmitted all along the epithelium. The excitatory state is not propagated by the movement of the cilia, and this movement can be stopped by cooling a stretch of epithelium without stopping the wave, which is transmitted through the cooled cells and excites the cilia of those on the further side of the block.

Tropisms. In unicellular organisms and in plants there is a particular form of movement known as tropism. An external physical force (physiotropism) or chemical substance (chemotropism) attracts (positive tropism) or repels (negative tropism) the organism or part of it. For instance, oxygen attracts some bacteria and repels others; similarly, light attracts the branches of green plants (phototropism) and gravity the roots (geotropism). Heat (thermotropism), mechanical contact (thigmotactism), fluid currents (rheotropism), or electricity (galvanotropism) also provoke this type of movement.

MUSCULAR CONTRACTION

Fundamental properties of muscle. In higher organisms there are highly differentiated, elongated cells, which have the property of diminishing one diameter and increasing the others, *i.e.*, which can shorten and thicken. These cells are the muscle fibers, and this type of movement is called "contraction"—improperly so, because only insignificant changes in volume occur during contraction and relaxation. The extremities of the fibers are attached to other structures (*e.g.*, bones), either directly or by means of fibrous formations (tendons), and when the muscle contracts these structures are moved. Muscles also form part of the walls of cavities or tubes, the diameters and capacity of which diminish on contraction of the wall muscles and increase on their relaxation.

There are three types of muscle fibers: striated, smooth or plain, and cardiac. They all have the same fundamental properties: (*a*) the capacity of responding to a stimulus, *excitability;* (*b*) the capacity to transmit the excitatory state, *conductivity;* (*c*) the capacity to shorten, *contractility;* (*d*) the capacity to recover the resting form after a force has ceased to act on them, *elasticity;* (*e*) resistance to change in shape due to internal friction, *viscosity.* Each type of muscle has distinctive structural and functional characteristics. Striated muscle is also known as "skeletal," "somatic," or "voluntary" muscle, because in vertebrates it is inserted on the bones, is originated in the somites, and is usually brought into activity by impulses from the central nervous system. Smooth muscle is sometimes called "visceral," because it forms part of the walls of viscera, or "involuntary," because it has a considerable degree of autonomy. Striated muscle will be considered first.

Structural Aspects of Striated Muscle

Microscopic structure of muscle fibers. Muscle fibers are giant cells, formed in the course of development by fusion of several mesenchymatic cells, the myoblasts. They are enclosed in a thin structureless membrane, the *sarcolemma.* Histochemical examination has shown that this membrane has no collagenous substance; therefore it is probably formed exclusively by the myoblasts, without the participation of connective-tissue cells. Microincineration shows that it is rich in silicates. The myofibrils are surrounded by a network of filaments, visible with the electron microscope, which have a tendency to unite and form continuous sheets. Each filament is about 250 A in diameter and is made up of a series of disks some 200 A high. They seem to be related to collagen and probably contribute greatly to the stability and the elastic properties of muscle.[1]

The *nucleus* of the myoblast divides several times, but the cell does not divide; thus the adult fiber has several nuclei, situated mainly under the sarcolemma. The *protoplasm* is made up

[1] Rosza, G., A. Szent-Györgyi, and R. W. G. Wickoff, *Exper. Cell Research,* 1, 194, 1950.

of two different parts, the *fibrils* and the *sarcoplasm*, which is found between the fibrils and under the sarcolemma. In the neighborhood of the nuclei, a Golgi apparatus, mitochondria, and lipid inclusions are seen; these lipid bodies are also found between the fibrils. The myoblast forms a fibril which later divides into bundles of fibrils, the *sarcostyles*. The outer layer of the sarcostyles behaves as a semipermeable membrane.

The sarcostyles are made up of segments, the *sarcomeres*. (Fig. 339, *S*.) A relatively thick, resistant, and impermeable membrane separates one sarcomere from another; it is called Krause's membrane, or the Z disk. Krause's membranes of neighboring sarcostyles are joined together and finally inserted in the sarcolemma; thus they form partitions which hold at the same level all the sarcostyles of a muscle fiber.[1] This division of the sarcostyles into segments gives the fiber the cross-striated aspect particular to this type of muscle. Observations of unstained myofibrils made with the electron microscope have shown transverse bands corresponding to this cross striation. Longitudinal striation, due to the fibrillar constitution, is common to all types of muscle.

The middle part of the sarcomere, which is dark, is known as Brücke's or the Q or A disk; it is separated from Krause's membrane by a clear, hyaline substance, called the J or I disk. Examined under the microscope by polarized light, the Q or A disks are seen to be anisotropous, but the J disk is only slightly doubly refracting; its anisotropy is about 10 per cent that of the Q disk. The ratio of Q to J in a resting fiber is 55:45. Q is rich in salts, especially in potassium chloride and phosphates; Fe is also found. Microincineration shows J to be almost free from minerals. A clear disk, the line of Hensen (*H*), divides Q into two equal parts. In a stretched fiber a fine membrane, Heidenhain's membrane (*M*), can be seen in the middle of *H*. The J disk shows a dark line about halfway between the membrane of Krause (*Z*) and the Q disk; this is Engelmann's accessory disk (*N*), which probably is an artefact. There is also another dark line close to *Z* known as the terminal disk.

[1] Doubts have been cast on the existence of Krause's membrane in the living fiber. Coupling of the contractile filaments and the myofibrils with each other and the sarcolemma has been attributed to physical forces.

Observations made with the electron microscope[1] have shown that the myofibril is made up of closely packed filaments[2] which can be separated by mechanical means in a colloid mill, or by tryptic digestion. They run the whole length of the sarcomere but do not cross the Z membrane. They are about 200 A in diameter and have nodosities or differences in densities at regular intervals of approximately 400 A. In the intact myofibril "the particles are arranged with great regularity standing in rows both cross- and length-wise, that is, associated both end to end and side to side" (Szent-Györgyi), giving the myofibril a fine cross-banding distinct from the microscopic cross striation. This tridimensional regularity in structure is not peculiar to muscle fibers, but is also found in other animal and vegetable fibers. The spaces between the different layers and particles are not uniform in the three dimensions of space, and they are small enough with respect to the wavelength of light to endow the system with the property of double

FIG. 339. Electron-microscope images of skeletal muscle fibers of toad (×20,000). (*Draper, M. H., and A. J. Hodge, Nature, vol. 163, p. 576, 1949.*)

[1] HALL, C. E., M. A. JAKUS, and F. O. SCHMITT, *Biol. Bull.*, **90**, 32, 1946; MORGAN, C., G. ROZSA, A. SZENT-GYÖRGYI, and R. W. G. WYKOFF, *Science*, **111**, 201, 1950; ASHLEY, C. A., K. R. PORTER, D. E. PHILPOTT, and G. M. HAAS, *J. Exper. Med.*, **94**, 9, 1951.

[2] In the living myofibril there are no preformed filaments; the contractile substance forms a continuous mass. The filaments appear owing to the treatment to which the muscle is submitted.

refraction. Muscle fibers are birefringent or anisotropous (Fig. 340).

Evidence obtained with the electron microscope suggests that the myofibril has a tubelike structure, the walls containing the filaments and an A substance, surrounding an aqueous core.[1]

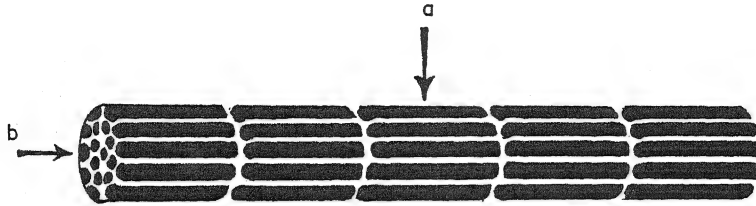



Fig. 340. Diagram of an anisotropous fiber. Observation along the axis *a*, or any other transverse axis, will show a series of layers placed at equal distances from each other. Observation along the longitudinal axis *b* will show a series of segments with identical characteristics at regular intervals; the spacing is not the same as along a transverse axis.

During contraction the A substance in the Q disks shifts toward the Z membrane and the M band is widened, becoming the most transparent part of the fibril. The dense substance on each side of the Z membrane forms the so-called "contraction band." The sarcomere shortens to half the resting length and becomes much wider.

There are two types of striped muscle fibers, the *white* and the *red* fibers. White fibers are thick, poor in sarcoplasm, with nuclei exclusively under the sarcolemma and in large numbers near the end-plate; cross striation is regular and well marked, and longitudinal striation is not easily distinguished. The red fibers have abundant sarcoplasm, with nuclei distributed all through the fiber; cross striation is irregular and poorly marked, and longitudinal striation is outstanding. Red fibers have more myohemoglobin and less cytochrome than white fibers. There are intermediate fibers between these two types. In some animals there are muscles made up mostly of either one or the other type of fiber; in man both kinds are found in all muscles, but in certain muscles one or the other type predominates.

The Excitation of Muscle

Muscles respond to many kinds of stimuli according to the general laws of excitability. Contraction can be provoked experimentally by mechanical, thermal, chemical, and electrical stimuli, but in physiologic conditions striated muscles are excited exclusively by nerve impulses.

Innervation of muscle fibers.[1] The nerves of skeletal muscles are made up of myelinated and unmyelinated afferent and efferent fibers. The different types of fiber can be separated, in order to examine their morphologic and physiologic characteristics, by (*a*) section of the dorsal roots (deafferentation); (*b*) section of the ventral roots (elimination of somatic motor fibers); (*c*) removal of the appropriate sympathetic ganglia (suppression of visceral fibers).

There are two types of motor fibers in the limb-muscle nerves of mammals: (*a*) large fibers, and (*b*) small fibers.[2] Large fibers in the cat have diameters between 10 and 18 μ (the majority between 12 and 14 μ), a low threshold, and high velocity of conduction (around 70 m./sec. and up to 108 m./sec.). Small fibers have diameters between 3 and 8 μ (peak at 5 μ) and a high threshold, and they conduct at velocities of about 27 and 30 m./sec. (not above 50 to 55 m./sec.).[3] Certain muscles (face, larynx, infrahyoid region) receive only medium-sized fibers distributed around a single peak at about 10 μ; these muscles have few spindles and perform relatively simple actions. Muscles in the limbs and the extrinsic eye muscles have small and large fibers distributed around two peaks; these muscles have many spindles, and they take part in complex accurate movements, or in postural reflexes.[4] In man there are many small

[1] Pease, D. C., and R. F. Baker, *Am. J. Anat.*, **84**, 175, 1949; Ashley, Porter, Philpott, and Haas, *loc. cit.*

[1] Hinsey, J. C., *Physiol. Rev.*, **14**, 514, 1934; Barker D., *Quart. J. Micr. Sc.*, **89**, 143, 1948; Tiegs, O. W., *Physiol. Rev.*, **33**, 90, 1953.

[2] Eccles, J. C., and C. S. Sherrington, *Proc. Roy. Soc., London, s.B.*, **106**, 326, 1930.

[3] Leksell, L., *Acta physiol. Scandinav.*, **10**, Suppl. 31, 1945; Kuffler, S. W., *et al.*, *J. Neurophysiol.*, **14**, 29, 1951.

[4] Fernand, V. S. V., and J. Z. Young, *Proc. Roy. Soc., London, s.B.*, **139**, 38, 1951.

fibers in the cervical roots but fewer in the lumbar and sacral roots.

In the frog, axons of small diameter (5 μ) innervate the same muscle fibers as the larger motor axons. Stimulation of the small fibers evokes local potentials similar to the end-plate potential (see page 793) and slowly rising, well-maintained local contractions, clearly distinguishable from the propagated contractions which involve the whole muscle fiber. With repetitive stimulation, tension up to 15 per cent of the maximal tension may be obtained. Small fibers can be stimulated reflexly, especially by stimulation of large afferent fibers, and the appreciable tension thus evoked can be maintained for prolonged periods. Impulses transmitted by the small fibers have been considered as playing a part in the maintenance of muscle tonus (see Chap. 82). Stimulation of the small fibers in the ventral roots of mammals, however, does not produce any appreciable increase in tension.[1]

In limb-muscle nerves (cat) large fibers make up two-thirds and small fibers one-third of the total efferent fibers. Large fibers pass into the muscle end-plate, which suffers degenerative changes and eventually atrophies when these fibers are cut. In vertebrates a few muscle fibers receive two nerve endings, which sometimes come from different spinal roots. Double innervation is the rule in invertebrates; usually one nerve fiber stimulates and the other inhibits the muscle fiber, as also occurs in most vertebrate smooth muscles. The small fibers innervate the intrafusal muscle fibers of the spindles.[2] The functional significance of the small fiber system will be discussed in Chap. 73 ("Proprioceptive sensibility") and Chap. 81 ("Muscle tonus").

In the frog the large efferent fibers not only innervate muscle fibers, but also give off branches to the intrafusal muscle fibers of most spindles.[3] The small nerve fibers innervate skeletal muscle fibers which have distinctive anatomical, physiologic, and pharmacologic characteristics.[4]

Afferent fibers in limb muscles have been classified into three groups in descending order

[1] Tasaki, I., and K. Mizutani, *Japan. J. M. Sc.*, **10**, 237, 1944; Kuffler, S. W., *Proc. Soc. Exper. Biol. & Med.*, **63**, 21, 1946; Kuffler, S. W., and Y. Laporte, *Federation Proc.*, **6**, 146, 1947.

[2] Kuffler, S. W., *Proc. Soc. Exper. Biol. & Med.*, **63**, 21, 1946.

[3] Katz, B., *J. Exper. Biol.*, **26**, 201, 1949.

[4] Kuffler, S. W., *Arch. d. sc. physiol.*, **3**, 613, 1949.

of diameter: groups I, II, and III.[1] The large group I fibers constitute 45 per cent of the afferent fibers of the hind-limb extensors and a smaller proportion of the flexors; they innervate the muscle spindles and Golgi tendon organs. Stimulation of the smaller fibers in groups II and III gives rise to the flexor reflex (see Chap. 69).

In the neighborhood of the muscle fiber the axon loses the myelin sheath, penetrates the sarcolemma, and ramifies in the sarcoplasm, which is abundant and has numerous nuclei in this place, known as the end-plate.

Section of the motor nerve is rapidly followed by degenerative changes and eventually by atrophy of the end-plate.

Nerves to the muscles include nonmyelinated fibers; some are afferent fibers, but the majority are postganglionic sympathetic fibers which innervate the blood vessels. No satisfactory proof has been given that the muscle fiber receives sympathetic as well as somatic innervation.[2]

When a muscle is stimulated, even after its nerve has been cut, the stimulus acts on the nerve fibers and only indirectly through these does it act on the muscle. It is therefore necessary to give time for nerve degeneration to take place if a purely muscular response is desired. Another method consists in the study of individual muscle fibers isolated in Ringer's solution. In this condition the nerve and end-plate degenerate within a few hours, but the properties of the muscle fibers remain apparently unaltered for several days. These fibers give strength-duration curves of the α type (see below, "Neuromuscular block") at whatever point the fiber is stimulated; therefore the excitability of the fiber is uniform throughout its whole length.[3]

Transmission of Excitation from Nerve to Muscle

The excitatory state is transmitted from the nerve to the muscle, but not in the opposite direction from the muscle to the nerve. In the end-plate as in the synapse there is unidirectional or irreversible conduction. This fact has great interest, not only for the understanding of

[1] Rexed, B., and P. O. Therman, *J. Neurophysiol.*, **11**, 133, 1948; Lloyd, D. P. C., and H. T. Chang, *J. Neurophysiol.*, **11**, 199, 1948.

[2] Hinsey, *loc. cit.*

[3] Ramsey, R. W., and S. S. Street, *Biol. Symposia*, **3**, 9, 1943.

muscular excitation, but also because phenomena occurring at the end-plate may throw light on synaptic transmission of the excitatory state.

End-plate delay. When the nerve impulse reaches the end-plate, there is a delay before the muscle contracts. The delay lasts 2 msec. in the frog gastrocnemius, 0.8 msec. in the cat gastrocnemius, and a slightly shorter time in very rapid muscles, such as the eye muscles. During this period electrical and chemical processes occur, and the excitatory state is transmitted from the nerve to the muscle.

Table 87. Changes in Chronaxies of Muscle and Nerve

Factor causing change	*Muscle*	*Nerve*
Curare	+	0
Sparteine	+	0
Fatigue	+	0
Strychnine	0	−
Nicotine	−	0
Veratrine	−	0
Scopolamine	+	−

+ = increase; − = decrease; 0 = no change.

Neuromuscular block. The transmission of excitation from the nerve to the muscle fiber may be blocked. The muscle then responds to direct stimulation but not to stimulation through the nerve, although the excitability of the latter is unimpaired and it conducts impulses of normal amplitude at normal velocity. Neuromuscular block was first observed by Claude Bernard (1851) when studying the effects of curare (a poison extracted from several plants of the genus *Strychnos* by Amazon Indians, who smear it on their arrows); hence the name "curarization" given to this phenomenon.

The peripheral action of curare was demonstrated by Bernard in the following experiment: A ligature was placed on the leg of a frog, leaving out the sciatic nerve, so that the circulation was suppressed without disturbing nerve connections with the centers. Curare was then injected into the animal. When paralysis was fully developed, stimulation of the sciatic nerve of the leg with normal circulation, *i.e.*, that on which curare had acted, did not provoke contraction, although the muscle responded to direct stimulation. Stimulation of the sciatic nerve of the leg in which the circulation had been suppressed, so that curare could not act on it, provoked contraction.

Curare kills by paralyzing the respiratory muscles. Death can be prevented by keeping up artificial respiration until the effect of the drug has passed off. Several alkaloids have been extracted from curare, *e.g.*, curarine and D-tubocurarine. In surgical operations when muscular relaxation is required, it can be provoked without submitting the patient to deep anesthesia by injecting a curare alkaloid, taking care to maintain adequate pulmonary ventilation.

Kühne supposed that curare acted on a receptive substance in the end-plate, which serves as a bridge between the nerve and muscle fibers. The hypothesis of the existence of a neuromuscular junctional tissue received further support from observations by Langley on the action of nicotine on striated muscle and by Elliott on the action of adrenaline on smooth muscle. Strength-duration curves, however, show only two types of excitability,[1] known as α and γ. The former is slower; it is the only one found in nerve-free parts of muscle (sartorius) and in fully curarized muscles. The latter is faster, has the same time scale as the strength-duration curve of the nerve, and is abolished by curare; it is attributed to excitation of the nerve. There is no satisfactory evidence of a third type (β) of excitability which would be even faster and correspond to the junctional tissue.

Lapicque[2] has maintained that normally the chronaxie of muscle is subordinated to that of its nerve; there is isochronism, *i.e.*, muscle and nerve chronaxies have the same value. When either muscle or nerve chronaxie is altered so that they are in a 1:2 ratio (heterochronism) the neuromuscular junction is blocked. Neuromuscular block, however, can be produced by doses of curare which do not change the excitability of muscle, much larger doses being necessary to depress muscular excitability.[3]

Neuromuscular block is observed in transmission fatigue and can be produced by several drugs (Table 87), though not necessarily by the same mechanism as that of curare and its alkaloids; *e.g.*, acetylcholine; cholinesterase inactivators such as eserine, prostigmine, diisopropylfluorophosphate (DFP), etc.; decamethonium; erythrine, an alkaloid extracted

[1] Lucas, K., *J. Physiol.*, **36**, 115, 1907; Rushton, W. A. H., *J. Physiol.*, **70**, 318, 1930; **72**, 265, 1931; **74**, 231, 1932; **75**, 161 and 445, 1932.

[2] Lapicque, L., L'excitabilité en fonction du temps," Presses Universitaires de France, Paris, 1926; Lapicque, L., *J. Physiol.*, **73**, 189 and 219, 1931.

[3] Rosenblueth, A., *Am. J. Physiol.*, **129**, 22, 1940.

from the seeds of the "ceibo" tree, *Erythrina crista galli;* etc.

Curare acts by antagonizing the effect of the chemical mediator, acetylcholine, although it does not prevent the release of acetylcholine at the end-plate on arrival of the nerve impulse in the body, an *action potential* develops in which two components can be separated, the end-plate potential and the spike potential (propagated potential).

End-plate potential. Normally the end-plate potential (e.p.p.) is masked by the spike potential, but it can be made evident by means of curare or its alkaloids.[1] The effect of this drug is to reduce the amplitude of the e.p.p. below the threshold of the propagated impulse so that the spike potential does not appear. The e.p.p. consists of a single monophasic wave (Fig. 341) which rises to a peak in 1 to 1.5 msec. (completely curarized frog's sartorius at 20°C.) and decays gradually, taking 2.4 to 6.0 msec. from onset to half decline.[2] The amplitude in a fully curarized muscle is 10 to 20 mv., but it is less if the dose of curare is increased, and conversely the removal of curare is accompanied by an in-

S
10 mv.
0
0.5
1
1.5
2
2.5
3
3.5
4
0 10 20 30
A
10 mv.
0 10 20 30
B
50 mv.
0 1 2 3 4 5
C

FIG. 341. End-plate potentials. *A*, end-plate potential in single curarized muscle fiber, at the end-plate focus and at different distances up to 4 mm. from the focus. *B*, end-plate potential of curarized muscle before (above) and after (below) addition of prostigmine. *C*, spike potential at the end-plate (above) showing step (e.p.p.) in the ascending limb and "hump" in the descending limb; spike potential (below) at 2.5 mm. from the end-plate focus. Time in msec. (*Fatt, P., and B. Katz, J. Physiol., vol. 115, p. 320, 1951.*)

(see Chemical phenomena of neuromuscular transmission page 794).

Electrical phenomena of muscular activity. Muscle fibers, like other living cells, have polarized membranes, and activity is accompanied by electrical phenomena. Knowledge of these phenomena has advanced considerably since it has been possible to stimulate and register electrical events in single fibers, either isolated or *in situ*.

The *resting potential* of the muscle fiber of the frog at 20°C., measured with an intracellular electrode, is about 90 (75 to 105) mv., and there is not more than 0.5 mv. difference from one part of the fiber to another. Curare does not modify the resting potential. When the fiber is stimulated through its nerve, as occurs normally

[1] ECCLES, J. C., B. KATZ, S. W. KUFFLER, *J. Neurophysiol.*, 4, 362, 1941.

[2] FATT, P., and B. KATZ, *J. Physiol.*, **115**, 320, 1951.

crease in the rate of rise and amplitude until a spike is fired off. The e.p.p. spreads for a few millimeters along the fiber, decaying exponentially with time and distance from its focus in the end-plate. The e.p.p. increases with the strength of the stimulus, and by repeated stimulation at adequate intervals so that summation may take place, it may reach the threshold and a spike is fired off.

The e.p.p. is due to transient subliminal depolarization of the end-plate surface. In normal muscle during transmission of an impulse the e.p.p. is seen as an initial step of about 40 mv. occurring at the first ½ msec. of the rising phase of the action potential registered at the end-plate. A few millimeters from the end-plate the spike potential does not show this step (Fig. 341*C*).

Antagonists of cholinesterase, *e.g.*, prostigmine, increase the amplitude and duration of e.p.p. (Fig. 341*B*). Sodium and calcium deficiencies, on the contrary, reduce the e.p.p.

Minute end-plate potentials are registered in resting fibers; their amplitude is about 1 (0.4 to 2) mv., and they appear irregularly at random with occasional bursts of high frequency. They are reduced in size by small doses of curarine and increased by prostigmine. Denervation and nerve anesthetics abolish them.[1]

Spike potential. If depolarization of the membrane attains the threshold, a propagated disturbance is initiated and spreads over the surface of the whole fiber. It is accompanied by a spike potential with a reversal of about 35 mv. at its peak, which it reaches in 1 msec. It then decays and is at half-decline in approximately 4 msec. This is an "all-or-nothing" phenomenon, *i.e.*, it is not proportional to the strength of the stimulus; a threshold stimulus evokes the maximal potential.

In a normal muscle the action potential evoked by a nerve impulse and registered at the end-plate differs from the action potential registered at other parts of the muscle fiber and from that evoked by direct stimulation; its peak is lower by about 15 mv., and the descending limb does not fall smoothly but forms a "hump" attributed to the persistent action of the mediator (Fig. 341*C*).

The spike is followed by a refractory period. The velocity of spread is 2 m./sec. in the largest fibers of the frog at 20°C.; it is slowed down by cooling; the temperature coefficient Q_{10} is 2. In mammalian muscle the following figures have been found for speed of conduction in meters per second: soleus of cat, 2.85; tibialis anticus of cat, 5 to 6;[1] anterior gracilis of rat, 3.4 to 5.[2]

[1] Burn, B. D., and W. D. M. Paton, *J. Physiol.*, **115**, 41, 1951.

The action potential wave is followed after a short interval by a wave of contraction, which spreads at the same velocity as the spike. The active state on the surface is propagated inward throughout the cross section of the fiber, but the nature of the process which links excitation to contraction is as yet unknown. The active contractile state is set up very abruptly (see page 798), so it cannot be due to diffusion of a substance from the surface; it is a physical or physicochemical process released at the surface and then propagated inward.[3]

After-potentials. Before the spike has come to an end a second negative potential begins, known as the *first negative afterpotential;* it has a low voltage (5 mv.) and lasts 50 to 150 msec. A *second negative afterpotential* of about 0.2 mv., lasting up to 1 sec., has been reported. There is inconclusive evidence for the existence of a *positive afterpotential* appearing between the two negative afterpotentials, of approximately 1 mv. and lasting 250 msec.

Chemical phenomena of neuromuscular transmission. Stimulation of the motor nerve, even after section and degeneration of afferent and sympathetic fibers, causes the release of acetylcholine (ACh) at the end-plate,[4] in sufficient amounts to provoke contraction. Curare, in doses which provoke neuromuscular block, does not prevent the liberation of ACh. Acetylcholine is rapidly destroyed by cholinesterase, which is found in the end-plate—most of it within the refractory period.[5] Its effect, therefore, cannot normally spread or persist beyond a very short time. The output of ACh from active motor nerve endings depends upon the concentration of Na^+ in the external fluid. The reduction in the e.p.p. and neuromuscular block

[1] Eccles, J. C., and W. J. O'Connor, *J. Physiol.*, **97**, 44, 1939.

[2] Jarcho, L. W., *et al.*, *Am. J. Physiol.*, **168**, 446, 1952.

[3] Hill, A. V., *Proc. Roy. Soc., London, s.B.*, **136**, 399, 1950.

[4] Dale, H. H., and W. Feldberg, *J. Physiol.*, **81**, 39, 1934; Dale, H. H., W. Feldberg, and M. Vogt, *J. Physiol.*, **86**, 353, 1936.

[5] Marnay, A., and D. Nachmansohn, *J. Physiol.*, **92**, 37, 1938; Nachmansohn, D., *J. Physiol.* **95**, 29, 1939.

brought about by lack of sodium or calcium is due mainly to a reduction in the output of ACh. It has been suggested that at the nerve ending ACh, rather than K^+, is exchanged for Na^+ that enters the nerve fiber during the rising phase of activity.[1]

A small dose (2 to 20 μg) of ACh injected into the artery of a muscle provokes muscular contraction with a propagated spike potential similar to that following stimulation of the motor nerve; it can also provoke contracture, *i.e.*, a mechanical response without a propagated spike potential (see Contracture).[2] Substances with anticholinesterase activity such as eserine, prostigmine, and di-isopropylfluorophosphate (DFP) increase the excitatory effect of ACh. Denervation increases sensitiveness to acetylcholine; at first acetylcholine produces mainly contraction; later the response is mainly contracture.[3] Curare, on the contrary, antagonizes the effect and raises the threshold of ACh.

Acetylcholine has a specific depolarizing effect on the end-plate, which it does not exert on other parts of the muscle fiber. Acetylcholine depolarizes even when there is no sodium in the external fluid; probably it produces a large nonselective increase in permeability and an ion sink is established at the end-plate, through which the fiber membrane becomes depolarized.[4]

Large doses of acetylcholine depress the response to direct electrical stimulation of the muscle, and to stimulation of the motor nerve; complete neuromuscular block may be produced. This effect is potentiated by cholinesterase antagonists.

Substances with anticholinesterase activity (eserine, prostigmine, DFP) may provoke muscular contraction. They augment the effect of nerve stimulation of low frequency, provoking a repetitive response to single shocks,[5] but they have a depressive effect on stimulation at high frequencies.[6] The response to direct stimulation of the muscle is not modified by cholinesterase antagonists, therefore they exert their effect on the process of neuromuscular transmission.

[1] Fatt, P., and B. Katz, *J. Physiol.*, **118**, 73, 1952.

[2] Brown, G. L., *J. Physiol.*, **89**, 220, 1937; *Physiol. Rev.*, **17**, 485, 1937.

[3] Rosenblueth, A., and J. V. Luco, *Am. J. Physiol.*, **120**, 78, 1937

[4] Fatt, P., and B. Katz, *Proc. Roy. Soc., London, s.B.*, **140**, 183, 1952.

[5] Brown, G. D., H. H. Dale, and W. Feldberg, *J. Physiol.*, **87**, 394, 1936.

[6] Rosenblueth, A., *et al.*, *Am. J. Physiol.*, **115**, 53, 1936; *Science*, **84**, 551, 1936; Brown, G. L., B. D. Burns, and W. Feldberg, *J. Physiol.*, **106**, 36, 1947.

The strength of contraction decreases if the motor nerve is stimulated at a sufficiently high frequency (transmission fatigue); there is also a progressive decrease in the amount of acetylcholine liberated at the end-plate. In this condition eserine or another anticholinesterase increases the response of the muscle.

Potassium ions probably play a part in neuromuscular transmission. Potassium is released from muscles stimulated directly or through the motor nerve, or by acetylcholine. Curare prevents liberation of potassium by nerve stimulation.[1] In small doses potassium stimulates muscular contraction and can enhance the effects of motor-nerve stimulation and of acetylcholine. Large doses have a depressing effect. The stimulating action of potassium is not blocked by curare, but K ions have a decurarizing action. These facts have acquired special interest since the part played by potassium in excitation is becoming more fully understood (see Chap. 66).

The sequence of events taking place when a muscle is stimulated through its nerve can be interpreted as follows: The impulse arriving at the nerve endings does not cause a direct spread of electric current from nerve to muscle, but releases a mediator, acetylcholine, which raises the permeability of the end-plate membrane so that ions that do not diffuse through the resting membrane can now pass across it (depolarizing effect). The minute end-plate potentials of resting muscle are due to leakage of small amounts of acetylcholine. The depolarized area is the site of the e.p.p.; if this reaches the threshold (about 40 mv.), it stimulates the surrounding regions, local circuits are set up, and a propagated "all-or-nothing" spike potential spreads throughout the surface of the fiber. If the effect of acetylcholine is prolonged, repetitive firing will be produced as long as the e.p.p. stays above the threshold.

The end-plate can be blocked by (*a*) raising the threshold by strengthening polarization with anodal currents; (*b*) an antagonist of the mediator, such as curarine, which prevents depolarization; (*c*) persistent action of the mediator maintained by cholinesterase antagonists, *e.g.*,

[1] Fenn, W. O., *Proc. Soc. Exper. Biol. & Med.*, **37**, 71, 1937; Cicardo, V. H., *Rev. Soc. argent. de biol.*, **14**, 297, 1938; Cicardo, V. H., and J. L. Moglia, *Rev. Soc. argent. de biol.*, **16**, 149, 1940.

eserine, so that depolarization is abnormally prolonged and extended beyond the end-plate; (*d*) depolarization of the end-plate by cathodal currents.

Mechanical Phenomena of Muscular Contraction

Muscular contraction in the body is usually evoked by nerve impulses arriving at the end-plates of the fibers which make up the muscle. In order to analyze the complex phenomena which take place, muscles have been separated from the organism—in some cases, a single muscle fiber[1]—and stimulated directly by means of electric shocks. A single shock of sufficient strength, or a single nerve impulse, evokes a muscle twitch, *i.e.*, contraction followed by relaxation. This is the elementary unit of muscular activity. Repetitive stimulation of adequate frequency provokes sustained contraction called tetanus. If the resistance against which the muscle contracts is so great that it cannot shorten, tension develops and is converted into heat; this is known as an isometric contraction.[2] If the resistance is not so great, the muscle can shorten and perform work; this is known as an isotonic contraction.

A muscle twitch takes place very quickly; thus the contraction time (from the beginning of the spike potential to the peak of contraction) for the internal rectus of the eye is less than 8 msec. This can be slowed down by cooling, and at 0°C. the behavior of muscle is essentially the same as at normal body temperature, except

SE C SE

PE

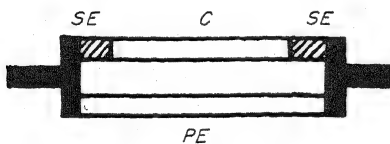



Fig. 342. Diagram of mechanical arrangement of muscle. *C*, contractile component; *PE*, parallel elastic component. (*Hill, A. V., Proc. Roy. Soc., London, s.B., vol. 138, p. 339, 1951.*)

for the lower rate of change. Slow muscles, such as those of the toad and tortoise, have also been used to obtain valuable information; thus skeletal muscles of the tortoise take fifteen times as long as those of the frog to complete contraction; *e.g.*, at 0°C. shortening begins 90 to 100 msec. after the stimulus has been applied, takes several seconds to reach the maximum, and takes several seconds more to complete relaxation.

[1] Ramsey and Street, *loc. cit.;* Høncke, P., *Acta physiol. Scandinav.*, **15**, Suppl. 48, 1947.

[2] Muscle is made to pull against a strong spring, the minute movements of which are amplified and recorded.

1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
Tension
3
2
1
2
0.6
0.8
1.0
1.2
1.4
Length

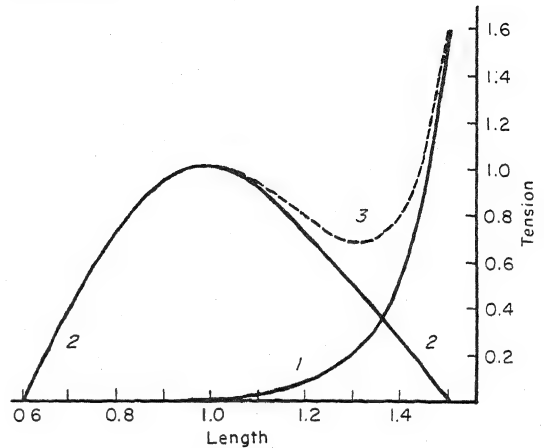



Fig. 343. Tension-length relations of sartorius of frog or toad. 1, at rest, positively stretched; 2, extra force developed during maximal tetanus; 3, total force of maximal tetanus (sum of 1 + 2). Length is given as fraction of standard length in body, and tension as fraction of maximum force developed (*Hill, A. V., Proc. Roy. Soc., London, s.B., vol. 141, p. 104, 1953.*)

Methods for registering the phenomena of muscular contraction must cope with their minuteness and their velocity; sensitiveness and speed are therefore essential. The use of electronic techniques, piezoelectric crystals, and ultimate recording with the cathode-ray oscillograph has made possible the registration of mechanical events with great sensitiveness and accuracy, because large amplification with insignificant inertia can be obtained.

Muscle is made up of (*a*) a contractile component, located in the A segment of the myofibril (see "Physico-chemical structure of the myofibril"); (*b*) a passive undamped elastic component in series with the contractile element, *e.g.*, tendons; (*c*) an elastic component in parallel with the two preceding components, constituted by the sarcolemma and fibroelastic tissues surrounding the muscle fibers (Fig. 342).

Elasticity. All the elastic components of muscle show the load-extension relation usually seen in biological fibers, *i.e.*, great extension with small loads which becomes less with greater ones (Fig. 343). The extensibility of the contractile component is so great that when a mus-

cle is stretched to little more than its resting length in the body, the load is borne mainly by the parallel elastic component. The behavior of the other two components (contractile and elastic in series) must be examined at lengths not greater than the normal resting length. When the muscle contracts in isometric conditions the series elastic component is stretched by the shortening of the contractile component. This extension is equivalent to 4 per cent or more of the length of the muscle at rest, depending on the length of the tendon. As the active state diminishes the contractile component becomes less capable of bearing a load or stress, and it is stretched by the series elastic component.

At rest the contractile component is not only extensible but also plastic. If a muscle is made to shorten by stimulation it remains shortened unless a load or elastic forces tend to pull or push it out. A resting muscle isolated from the body exerts a small but measurable tension at lengths down to about 60 to 75 per cent of the natural resting length in the body. If it is stretched to a given length, tension rises rapidly at first, then falls to a small value. Further stretching will again cause a rise in tension, followed by a fall to a small value. The resting muscle thus remains at any length it has been given. Its fibers are not folded but truly shortened, and if a quick stretch is applied, tension rises rapidly from the beginning; there is no delay, as there would be if slack of folded fibers had first to be taken up.

Resting muscle shortens when warmed under a constant load, or its tension rises if it is warmed while kept at constant length. Cooling produces the opposite effect. These are reversible processes and can be repeated any number of times. They are not due to excitation of the fibers, because they occur in muscles which have been made inexcitable by an excess of KCl. Muscle behaves, in this respect, like rubber. Other fibers, *e.g.*, tendon, have "normal" elasticity, *i.e.*, they lengthen on warming and shorten when cooled.

When a muscle is stretched there is an output of heat and a rise in temperature; on release heat is absorbed and the temperature falls. For a given distance of stretch, or release, the thermoelastic heat given out, or absorbed, becomes greater as the length is increased. These thermal changes are not simultaneous with the mechanical changes but show a considerable lag; they start slowly and continue after stretch or release is complete. At greater lengths thermoelastic temperature changes show a diphasic character; on release there is first an increase in temperature which appears without delay, followed by a fall in temperature. At short lengths there is only one elastic component involved, which has "rubberlike" elasticity, probably the contractile constituent of muscle; at greater lengths a second elastic component with "normal" elasticity comes into play, probably the parallel elastic component.[1]

The work done on a muscle by stretching it is much less than the heat produced, and the work done by a muscle released after stretching is less than the heat absorbed. The internal energy therefore decreases on lengthening and increases on shortening. This has been attributed to crystallization (an exothermic process) on stretching, as occurs in rubber. Changes in the birefringence of muscle suggest that crystallization increases on stretching. Stretching also increases the distance between the basic and acidic groups of the muscle protein chains; this increase would cause a shift to the alkaline side, as has in fact been observed,[2] and liberation of heat.

The stress-strain relations of a contracting muscle differ from those of a resting muscle. In a muscle in tetanic contraction tension developed by the active state is greatest at the length of the muscle at rest in the body, when the resting tension is very small. At greater lengths the resting tension, due mainly to the parallel elastic component, comes into play. The tension developed by contraction, obtained by subtracting the tension at rest from the total tension of the contracted muscle, diminishes as length increases (Fig. 343).[3]

Meyer[4] has postulated a model system which exhibits a stress-strain curve similar to the peculiar curve of contracting muscle. It consists of two spheres charged with opposite electric charges, which therefore attract each other, and rubber cylinders which when compressed compensate the attraction. Perhaps the charges in the basic and acidic groups of muscle

[1] Hill, A. V., *Proc. Roy. Soc., London, s.B.*, **139**, 464, 1952.

[2] Dubuisson, M., *Arch. internat. de physiologie*, **50**, 203, 1940.

[3] Hill, A. V., *Nature, London*, **166**, 415, 1950.

[4] Meyer, K. H., *Proc. Roy. Soc., London, s.B.*, **139**, 498, 1952.

proteins, the attraction and repulsion forces of which vary with shifts in pH, play a part similar to that of the spheres in Meyer's model.

The latent period. After a stimulus has been applied to a muscle there is a delay before the beginning of shortening or the rise in tension. This interval is known as the latent period.

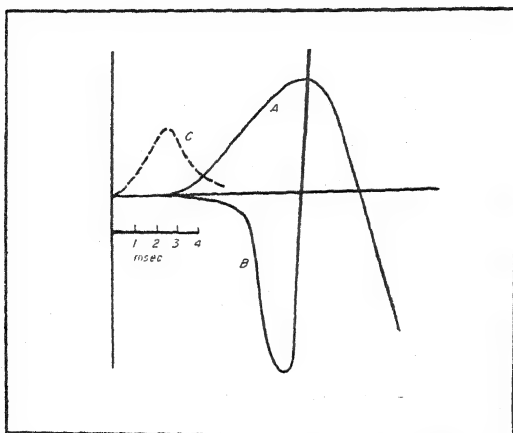



Fig. 344. Events in muscle following stimulation: *A*, opacity changes; *B*, mechanical response; *C*, course of action potential. (*Hill, D. K., J. Physiol., vol. 108, p. 292, 1949.*)

If muscle is not stimulated directly but through its nerve, part of the apparent latency is taken up by the spread of excitation along the axons (conduction time) and by the end-plate delay. "True latency," *i.e.*, that of the muscle itself, begins with the onset of the spike potential. When experimenting with a multifibered muscle, delay due to spread of excitation can be suppressed by using multiple electrodes so that all the fibers are stimulated simultaneously.

The latent period differs in different muscles; "quick" muscles have a shorter latency than "slow" muscles. Cooling prolongs the latent period and warming shortens it; the Q_{10} is about 2.1. In mammalian muscle at 37°C. it is of the order of 1 msec.

Several events occur in the course of the latent period:

1. About halfway in the latent period there is a slight lengthening or decrease in tension (Fig. 344). This initial or *latency relaxation*[1] is not observed if the muscle is not under a small initial tension. If the initial tension is decreased, latency relaxation diminishes and the phase of positive tension begins earlier. Hill[1] attributes latency relaxation to lengthening of the parallel elastic component.
2. There is an early increase in transparency, which begins at the same time as latency relaxation[2] (Fig. 344) and decreases as the muscle shortens. This is different from the increase in transparency which occurs later and is not complete until recovery has restored the muscle to its initial state (von Muralt). Changes in transparency are attributed to changes in the molecular pattern of the muscle proteins.
3. Early in the latent period, even before the onset of latency relaxation, the internal mechanical condition of the muscle changes abruptly. If it is submitted to a quick stretch, it is seen that the resistance to stretch is greater than that of resting muscle. The contractile component suddenly becomes capable of bearing a load equivalent to the maximal tension developed in an isometric contraction.[3] Resistance to stretch increases rapidly, reaching its maximum in about 1½ times the latent period; it remains at this level for a time, then diminishes gradually as relaxation sets in.
4. About halfway in the latent period, at the same moment that latency relaxation commences, an output of heat begins (see "Energetics of muscular contraction," "Heat of activation").

Shortening. At the end of the latent period the muscle begins to shorten. Speed of shortening increases rapidly and reaches its full value in about 1½ times the latent period (7 per cent of the time to maximum shortening); it remains constant in rate for a time, then falls off and is followed by lengthening during relaxation if the muscle has contracted against a load.[4] The time taken to reach full velocity of shortening is due in part to the fact that not all the fibers enter into the active state simultaneously. The active state which causes shortening, therefore, is fully developed in each individual fiber in a very

[1] Rauh, F., *Ztschr. f. Biol.*, **76**, 25, 1922; Sandow, A., *Ann. New York Acad. Sc.*, **47**, 895, 1947; Abbott, B. C., and J. M. Ritche, *J. Physiol.*, **113**, 330, 1951.

[1] Hill, A. V., *Proc. Roy. Soc., London, s.B.*, **138**, 339, 1951.
[2] Hill, D. K., *J. Physiol.*, **108**, 292, 1949.
[3] Hill, A. V., *Proc. Roy. Soc., London, s.B.*, **136**, 399, 1950; **138**, 399, 1951.
[4] *Ibid.*, **138**, 329, 1951.

short time, not more than one-tenth the time taken to develop maximum tension.

Speed of shortening against a load diminishes as the load increases (Fig. 345). The relation of speed to load can be expressed by the following equation:[1] $(P + a)v = b(P_0 - P)$, where P is the load, v is velocity of shortening, P_0 is the maximum force the muscle can exert at zero speed, and a and b are constants with the dimensions of force and velocity respectively. The maximum speed of shortening under zero load is about $4b$. The constant a is especially significant because during shortening there is an output of heat (see "Energetics of muscular contraction"), proportional to the amount of shortening, which is defined by ax (x is the distance shortened).

This characteristic relation between force and speed of shortening holds for skeletal muscle of frog, toad, and man,[2] and there is certainly a relation of this type in other muscles.

The force developed by contraction varies with the initial length of the muscle; it is greatest when the length is approximately that of the muscle at rest in the body.

If an isolated fiber of a frog muscle is allowed to shorten to less than 60 to 70 per cent of its resting length, its properties are permanently altered: it no longer lengthens in relaxation, but remains shortened; the tension developed on stimulation is less and the time for its full development is greater; there are structural alterations in the fiber, which loses the regularity of its cross striation. This is called by Ramsey and Street[3] the δ state. Similar structural changes are seen in contractures.

Relaxation. At different times relaxation has been considered as a passive or an active phenomenon. There is no heat of relaxation; whatever heat is released during the period of relaxation can be accounted for by degradation of mechanical energy (see "Energetics of muscular contraction"). Moreover, if a muscle shortens during contraction and it is not submitted to a load or tension, it does not lengthen but remains at the length to which shortening during contraction has reduced it. Relaxation seems, therefore, to be a purely passive process.[1]

[1] Hill, A. V., *Proc. Roy. Soc., London, s.B.*, **126**, 136, 1938; *Nature, London*, **166**, 415, 1950.

Ralston, H. J., V. T. Inman, L. A. Strait, and M. D. Shaffroth, *Am. J. Physiol.*, **151**, 612, 1947; Wilkie, D. R., *J. Physiol.*, **110**, 249, 1950; Bigland, B., and O. C. J. Lippold, *J. Physiol.*, **123**, 214, 1954.

[3] Ramsey and Street, *loc. cit.*

Changes in volume. The volume of a muscle diminishes rapidly as soon as contraction begins and continues to decrease as contraction goes on; with relaxation the volume increases suddenly, but there is a small remaining constriction even after a single twitch, which is larger after a tetanus. After relaxation the volume again increases. Reversible changes in volume accompanying contraction and relaxation are due to mechanical compression and decompression, chiefly of water contained in the muscle. Internal pressure rises to 100 to 300 mm. Hg[2] and stops the blood flow through the muscles. The decrease in volume remaining after relaxation is due to chemical breakdown, probably of adenosinetriphosphate and creatinephosphate. Later the volume again increases, corresponding to the resynthesis of the substances broken down; if resynthesis of phosphagen and the formation of lactic acid are prevented by poisoning with iodacetate, constriction may go on increasing after relaxation.

Fig. 345. Relation between force and velocity for human arm muscles. Dots, mean of 30 determinations of velocity. (*Wilkie, D. R., J. Physiol., vol. 110, p. 249, 1950.*)

When a muscle is stretched there is a small decrease in volume which is attributed to crystallization, as occurs in rubber.

Repetitive stimulation. Summation. A muscle preparation which has been rested for approximately an hour and which is then stimu-

[1] Hill, A. V., *Proc. Roy. Soc., London, s.B.*, **136**, 420, 1949.

[2] Hill, A. V., *J. Physiol.*, **107**, 518, 1948.

lated so as to obtain a maximal contraction, *i.e.*, contraction of all the muscle units, gives a first response that is less than the second. The second response in turn is not as great as the third, and so on, until after a few contractions their height and duration become stabilized. This was first

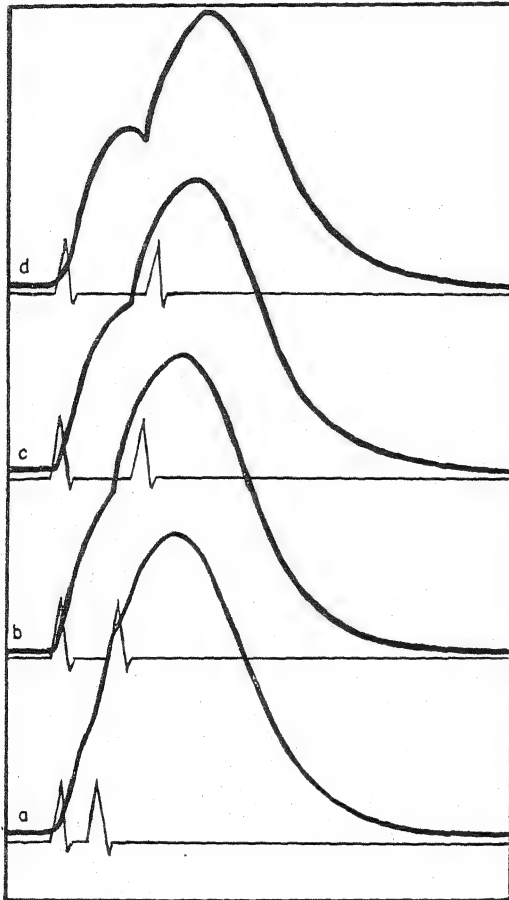



Fig. 346. Mechanical and electrical response of the cat's gastrocnemius to a double induction shock on the peroneal nerve. Interval between shocks: *a*, 24 msec.; *b*, 32 msec.; *c*, 40 msec.; *d*, 48 msec. (*Cooper and Eccles, J. Physiol., vol. 69, p. 377, 1930.*)

observed by Bowditch in the frog's heart, brought to a standstill by a Stannius ligature; he called it the "staircase effect," (*Treppe*) owing to the aspect of the record obtained.

When two maximal stimuli are sent with an interval such that the second arrives before the end of the effect produced by the first, the second contraction develops a greater tension and lasts longer than the contraction produced by a single stimulus. When the second stimulus falls during the absolute refractory period left by the first stimulus, it evokes no response. As the interval between the stimuli is lengthened, the response to the second stimulus becomes greater; it reaches a maximum when the second stimulus is applied at the peak of the response to the first stimulus (Fig. 346). The first stimulus provokes shortening of the contractile component and the series elastic component is stretched. The active state, however, commences to decline and relaxation to begin before stretching is complete. The second stimulus brings the active state again to its full intensity; the contractile component shortens and commences to pull on the series component which is already stretched, and a higher tension will be developed than can be achieved by a single shock. There is no algebraic summation of the effects of the consecutive stimuli; tension developed by each one will depend on the amount of stretch already existing. This is called "wave summation," because it is the sum of two waves of contraction in the same fiber.

When the muscle is stimulated repeatedly at a frequency such that the interval between stimuli is less than the contraction time, a sustained contraction of several times the strength of the simple contraction is observed. This type of contraction is called "tetanus." When the frequency of stimulation is not so high, relaxation commences before the succeeding stimulus arrives, the contraction is not so strong, and the top of the curve shows undulations, which are more marked as the frequency decreases. This type of contraction is called "incomplete tetanus" (Fig. 347). The minimum frequency of excitation needed to produce a complete tetanus varies inversely to the contraction time. Rapid muscles need a higher frequency than slow muscles, *e.g.*, this frequency is 30 per second for the soleus of the cat, 100 per second for the gastrocnemius, and 350 per second for the internal rectus of the eye, the contraction times of which are 100, 40, and 8 msec. respectively.

Initial tension or length (within certain limits) increases the maximum tension, and the tension developed in a tetanus, in the same way as in a muscle twitch. Once stimulation has ceased, the muscle remains contracted in tetanus for a short time, which is prolonged by increasing the initial length of the muscle and the duration and (within certain limits) frequency of stimulation.

After a brief tetanus, a single supramaximal shock or an injection of acetylcholine provokes a stronger twitch than before the tetanus. *Posttetanic potentiation* disappears gradually during the following rest period. It is observed in single isolated fibers and is not suppressed by complete curarization; therefore, it is not due to recruiting of fibers but to some change in the muscle fiber which favors a higher development of tension. In slow muscles, *e.g.*, the soleus, after a short high-frequency tetanus a single shock may provoke a repetitive response followed by a period of inhibition, which may last up to 100 msec., during which the response to a single shock is considerably diminished.[1] Curare suppresses this type of posttetanic potentiation; therefore it is due to a change in synaptic transmission. The demarcation potential diminishes after a brief tetanus. A similar decrease in potential and increase in response to a single shock is provoked by an injection of KCl, facts which have led to the suggestion that posttetanic potentiation is due to mobilization of potassium.[2]

The motor unit. The axon of a motor neuron branches out in its course toward the periphery and ends in a number of muscle fibers. An individual motor neuron together with the bunch of muscle fibers it innervates is a motor unit. When a nerve impulse is transmitted along the axon, all the muscle fibers at which its branches end are stimulated and contract simultaneously. The ratio of muscle fibers to nerve fibers gives the average number of muscle fibers per unit in a muscle, but the individual units are of variable size. The units in the limb muscles, which function in the maintenance of posture and locomotion, are mainly large units, although there are also smaller ones. Muscles that take part in finely graduated movements, such as the hand and eye muscles, are made up of small units. In the cat, the innervation ratio of the soleus, a red muscle, is 1:120, and that of the extensor digitorum longus, a white muscle, is 1:165. The eye muscles of man have innervation ratios of 1:2 and 1:6. The fibers of a unit are distributed along the whole length of the muscle, not grouped in one part of it, so that when the unit contracts tension is developed not at just one point but throughout the muscle, thus making for smooth contraction.

The average tension developed by the contraction of a motor unit is measured in the following way: The corresponding dorsal root ganglia are removed, and time is given for the complete degeneration of the sensory fibers. The motor nerve is then stimulated with a maximal stimulus, and the tension developed is measured.

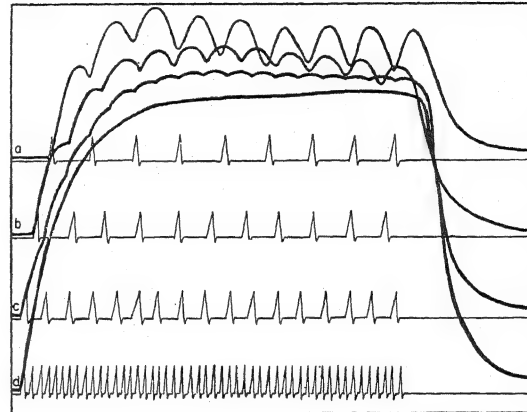



FIG. 347. Mechanical and electrical response of the cat's gastrocnemius to induction shocks on the popliteal nerve. Frequency of stimulation: *a*, 19 per second; *b*, 23.5 per second; *c*, 35 per second; *d*, 115 per second. This last frequency provokes complete tetanus. In *a*, *b*, and *c*, relaxation commences before the following stimulus is applied; oscillations in the mechanical record correspond to those of the electrical record. (*Cooper and Eccles, J. Physiol., vol. 69, p. 377, 1930.*)

Table 88. Number of Motor Units and Average Tension per Unit in the Cat's Muscles

Muscle	*Tension per unit, gm.*	*Number of units*
Gastrocnemius	30.1	430
Soleus	9.9	200
Semitendinosus	5.5	630
Extensor digitorum longus	8.6	330
Crureus	10.2	250

Source: ECCLES, J. C., and C. S. SHERRINGTON, *Proc. Roy. Soc., London, s.B.*, **106**, 326, 1930.

The fibers in the motor nerve are counted, and the maximum tension is divided by the number of fibers to obtain the average tension per motor unit (Table 88). Large units develop more tension than the average, and small ones less.

[1] FENG, T. P., T. H. LI, and Y. C. TING, *Chinese J. Physiol.*, **14**, 55, 1939.

[2] ROSENBLUETH, A., and R. S. MORISON, *Am. J. Physiol.*, **119**, 236, 1937; BROWN, G. L., and U. S. VON EULER, *J. Physiol.*, **93**, 39, 1938.

Muscular contraction in the organism. Adrian and Bronk[1] studied the motor nerve impulses to a muscle by means of a preparation in which the nerve was reduced to a small number of fibers and the action potentials were registered. Striated muscles normally receive nerve impulses of low frequency, which provoke a certain amount of tension, called muscle tonus. When a voluntary or reflex movement takes place, the frequency of the impulses increases and the muscles contract. The maximal frequencies observed were not sufficient to produce a complete tetanus; they were below 15 per second in the nerve of a slow muscle such as the soleus, and were not above 100 per second even with maximal stimulation in the most rapid muscles. Nevertheless a smooth contraction is obtained, because the units do not contract synchronously; there are fibers in different phases of contraction, and the oscillations of the incomplete tetanus are damped down.

FIG. 348. Duration of contraction and realizable work of forearm flexors in man. (*Adapted from A. V. Hill.*)

The strength of contraction is graded by two different mechanisms: (*a*) the frequency of the nerve impulses, which conditions wave summation and therefore tension developed by each fiber; (*b*) activation of a greater or smaller number of units (multifiber or quantal summation). A vigorous contraction throws into activity a large number, or even all, of the units of a muscle, at a high frequency of stimulation. There is also a tendency toward synchronization of the nerve volleys, and large waves appear in the electromyogram, due to the summation of the action potentials of many fibers activated simultaneously. When the frequency of the nerve volleys is not sufficient to produce a complete tetanus, the contraction becomes oscillating, as can be observed on making a great muscular effort.

[1] ADRIAN, A. D., and D. W. BRONK, *J. Physiol.*, **67**, 119, 1929.

The degree of activity, *i.e.*, the number of active motor units and their frequency of excitation, has been determined in intact human muscles by integrating the action potentials with surface electrodes. Electrical activity, thus measured, is directly proportional to the tension developed and to the speed of shortening.[1] The maximal tension that can be developed by voluntary contraction (arm flexors) remains constant over a period of months[2] and seems to be as great as the maximal tension developed by electrical stimulation.[3]

The amount of energy that can be made to yield work, *i.e.*, the realizable work, is conditioned by the speed of contraction. The force developed by contraction is a decreasing linear function of velocity; therefore maximum force is developed at zero speed. On the other hand, energy expended in maintaining the active state is greater the longer it lasts. Therefore, there is an optimum rate of contraction for each muscle or group of muscles; for the forearm flexors in man it is 2 sec. (Fig. 348).

Maximum work is obtained when (*a*) the speed of contraction is the optimum for the muscle groups in activity; (*b*) the initial length is the optimum for the development of force in the active state; (*c*) the muscles contract against a weight which is just sufficient to require the maximum development of energy, but not such that shortening will be hindered (otherwise the contraction will be in part isometric). Training raises efficiency because contraction starts at the most favorable initial length (adoption of adequate posture) and is carried out at optimum speed.

Fatigue. Sustained repetitive stimulation is followed after a time by a progressive decrease in the mechanical response of the muscle; after a period of rest the muscle recovers its capacity to respond. This may be due to failure of (*a*) trans-

[1] LIPPOLD, O. C. J., *J. Physiol.*, **117**, 492, 1952; BIGLAND, B., and O. C. J. LIPPOLD, *J. Physiol.*, **123**, 214, 1954.

[2] WILKIE, D. R., *J. Physiol.*, **110**, 249, 1950.

[3] MERTON, P. A., *J. Physiol.*, **123**, 553, 1954.

mission of excitation from the nerve to the muscle fiber, *transmission fatigue*, or (*b*) the contractile mechanism, *contraction fatigue*.[1]

Transmission fatigue is produced by prolonged stimulation of the motor nerve at a sufficiently high frequency (*e.g.*, 60 per second for 1 hr.). Direct stimulation shows there is no loss in the contractile capacity of the muscle, even when its response to nerve stimulation is considerably reduced. In this type of fatigue the acetylcholine output per impulse diminishes, and the response is improved by acetylcholine or cholinesterase antagonists. Small doses of curare, which raise the threshold to acetylcholine, have a greater depressor effect on fatigued muscle (with a low acetylcholine output) than on fresh muscle (with a normally high acetylcholine output). Nerve and muscle excitability diminish during fatigue because of an increase in the rheobase.

The excitability of the muscle fiber itself is also depressed by repetitive stimulation. This has been demonstrated in single isolated muscle after degeneration of the nerve in the following way: The fiber is stimulated at a frequency high enough to produce complete tetanus. After a short time, tension begins to fall off and incomplete tetanus (oscillating response) replaces complete tetanus. A little later only every other stimulus provokes a response, and as time passes the fiber responds to every third, fifth, tenth, twentieth stimulus. This is a Wedensky-inhibition phenomenon attributed to lengthening of the refractory period.[1]

If the frequency of stimulation is low (less than 30 per second), there will be no transmission fatigue and the progressive fall in the mechanical response will be due to contraction fatigue. Contraction time lengthens. Relaxation becomes slower and less complete as fatigue progresses (Fig. 349). The amplitude of the isotonic contraction diminishes, because shortening is progressively reduced and relaxation is incomplete. If the load is increased, fatigue develops more rapidly (Fig. 350). Conversion of chemical energy into tension becomes less efficient, and the $Tl:H$ ratio diminishes (T = tension, l = length, H = heat). Metabolic products, *e.g.*, lactic acid, accumulate and their elimination hastens recovery. Energy-producing substances, *e.g.*, glycogen, diminish, and fatigue occurs very rapidly if the energy stores have been depleted, or the recovery processes are inhibited, *e.g.*, in muscles intoxicated with moniodoacetate (see "Chemical phenomena of muscular contraction"). A completely fatigued muscle goes into contracture (fatigue cramp).

In maximal voluntary effort of the whole body, *e.g.*, in running, the supply of oxygen is the factor limiting performance (see Chap. 47); the onset of fatigue is due to failure of the contractile mechanism (contraction fatigue). Maximal voluntary contractions of a single small muscle (adductor of the thumb) are also limited by the blood supply; fatigue does not seem to be caused by failure of neuromuscular transmission be-

min 0 2 4 6

Fig. 349. Fatigue. Isotonic contractions of a muscle-nerve preparation of the gastrocnemius of the toad, stimulated by submaximal induction shocks at the rate of 85 per minute. At first the height of contraction gradually increases and then decreases. Relaxation is incomplete after a time, and after 3 min. 30 sec. there is marked contracture.

[1] Rosenblueth, A., "The Transmission of Nerve Impulses in Neuroeffector Junctions and Peripheral Synapses," Wiley, New York, 1950.

[1] Ramsey and Street, *loc. cit.*

cause there is no change in the action potentials evoked by maximal shocks delivered to the nerve.[1]

Contracture. In certain conditions the development of tension and shortening are kept up for a long time, although there is no tetanus, *i.e.*, no summation of the effects of repetitive stimulation. This particular form of contraction is known as contracture. It is a reversible process, and therefore differs from cadaveric rigidity and from rigidity due to coagulation of the muscle proteins. In contracture there is an increase in the metabolism of muscle, heat production is considerably greater than at rest, and there are glycolysis and lactic-acid formation. The fibers in contracture are electronegative with respect to the resting fibers. A contracture differs from normal contraction in two ways: (*a*) there is no propagated excitatory disturbance, with a corresponding spike potential, whether the whole fiber or only part of it is in contracture; (*b*) the intensity of contracture is graded in respect to the strength of the stimulus; it is not an all-or-nothing phenomenon.

Experimental contractures can be produced in isolated or intact muscles in many ways. Repetitive supramaximal stimulation produces a contraction followed by incomplete relaxation, the degree of contracture being dependent on the strength of the stimulus. After 100 stimuli have been sent at a frequency of 30 per minute, the contracture falls off. This is known as "Tiegel's contracture" and is due to a prolonged mobilization of energy following stimulation. Mosso observed it on stimulating the muscles of the forearm in man. It is not of reflex origin as it has been obtained after anesthetizing the brachial plexus.

Heat (45 to 50°C.), high pressure, and mechanical stimulation can produce contracture. Acids also provoke contracture, which rapidly passes on to an irreversible rigidity. A large number of drugs produce contractures (chloroform and other anesthetics, veratrine, aconitine, Ba, Ca, etc.). Other drugs, such as nicotine, acetylcholine, and quaternary ammonias, provoke contracture only in denervated muscles.

a

b

Fig. 350. Voluntary contractions of the flexor muscles of the middle finger of the right hand, registered by Mosso's ergograph. *a*, weight lifted, 4 kg.; the height of the contraction diminishes progressively; *b*, weight lifted, 2 kg.; the height of contraction diminishes at first and then remains at a constant level (steady state) during the 8 min. of the experiment.

[1] Merton, P. A., *J. Physiol.*, **123**, 553, 1954.

In most cases a so-called "contracture" observed in the intact organism is not a true contracture, but an increase in muscle tonus or a prolonged tetanus, due to an increase in the frequency of nerve volleys. In certain conditions it is possible to obtain a true contracture, called "idiomuscular contracture" by Schiff. Ranson[1] has described a "myostatic contracture" which is observed in cases of rupture of the muscle tendon, tetanus, and prolonged immobilization in a plaster cast. This is not a true contracture but is rigidity due to structural alterations in the affected muscles.

Chemical Phenomena of Muscular Contraction

Muscle is a machine for converting chemical energy into mechanical energy. Energy is stored in certain substances in the muscle, which on disintegrating release energy by processes called exothermic or exergonic reactions. Moreover, "in the process of contraction the muscle machine itself undergoes chemical and physical changes."[2] In order to understand how muscle functions, therefore, it is necessary to study (*a*) its physicochemical architecture, and (*b*) the metabolic processes of contraction.

Physicochemical structure of the myofibril. Proteins make up about 20 per cent of the muscle substance, and four-fifths of its solids; carbohydrates, lipids, organic crystalloids, and inorganic salts make up the rest.

There are several water-soluble proteins in muscle; the main ones are the *myogens* and *mvoalbumin*. The latter seems to play a part in the building up of other muscle proteins during embryonic development. The insoluble fraction contains those which are most important from the functional point of view; the principal ones are (*a*) myosin, (*b*) actin, and (*c*) actomyosin.

Myosin constitutes 40 per cent of the total protein and about 8 per cent of the muscle substance. Its molecular weight is 840,000. X-ray diffraction studies have shown that myosin molecules consist of long chains of polypeptides, joined to each other by side chains so as to form a grid, which in turn is joined to other grids by other side chains (Fig. 351). The distance between the grids is the same as that between the chains in a grid; therefore, along two axes the spaces between the components of the molecule are the same. In the longitudinal axis the spacing is different; the "links" of the chain are not of the same length as the space between grids and chains. The myosin molecule is thus a uniaxial optical system, therefore anisotropous or birefringent (crystalline birefringence)[1] (Fig. 340). The chains can be more or less extended. In a stretched myosin fiber (Astbury's β type) the length of one segment in the longitudinal axis is 10.2 A. When the fiber is not stretched it folds (Astbury's α type) and the segmental length is reduced to 5.1 A. In the resting muscle fibers myosin is probably in the α and partly in the β configuration. In contraction it folds even more, taking on Astbury's supercontracted state.

Fig. 351. Diagram of a myosin molecule. The principal chains are joined together by side chains forming grids.

Myosin particles, or micellae, are 23 A in thickness and 2,300 A in length.[2] In suitable dilution, myosin shows flow double refringence because the elongated particles are placed

[1] Ranson, S. W., *J. Neurol., Neurosurg. & Psychiat.*, **8**, 304, 1928.

[2] Dubuisson, M., *Proc. Roy. Soc., London, s.B.*, **137**, 63, 1950.

[1] Structures composed of units with tridimensional regularity are crystals. Myosin and other fibrous proteins are dispersed in a liquid medium; they are therefore called "liquid crystals."

[2] Weber, H. H., *Proc. Roy. Soc., London, s.B.*, **137**, 50, 1950. According to Mommaerts, the particle measures 30 by 1500 A.

lengthwise in the direction of flow and have a tendency to lateral association, forming swarms of micellae. Anisotropy of myosin is due, therefore, not only to its molecular structure (crystalline double refraction) but also to the orientation of its micellae (form double refraction). The electron microscope shows the formation of a network of long threads. Myosin is one of a family of fibrous proteins, which includes keratin, epidermin, fibrinogen, fibrin, etc.

Myosin acts as an enzyme, splitting off the terminal phosphate of adenosinetriphosphate (ATP).[1] Attempts to separate this adenosinetriphosphatase activity from myosin have been unsuccessful. It is dependent on the presence of free SH groups in myosin and is inhibited by Mg^{++} at pH 6.5 to 7. Actomyosin, however, is active as ATP-ase at pH 7, and both Mg^{++} and Ca^{++} promote this activity. Myosin, in the presence of salts, binds ATP; Mg^{++} enhances this activity. In resting muscle prevailing dissociation of COOH groups of its constituent amino acids gives myosin a negative charge, but adsorption of Mg^{++} neutralizes this charge; ATP adsorbed by myosin, however, gives it a negative charge, which increases as ATP is taken up and decreases as ATP diminishes.

Myosin also acts as a deaminase, splitting off the amino groups of adenosinemonophosphate (AMP). This enzymatic activity is independent of the ATP-ase activity.

Actin, isolated by Straub, forms 12 to 15 per cent of the total protein in muscle. It has a molecular weight of 70,000. It exists in two forms, globular G-actin and fibrous F-actin. One can be converted into the other by a reversible process of polymerization-depolymerization. G-F transformation does not take place in a salt-free medium. Hydrogen ion is the most potent activator of polymerization, which takes place readily at pH 6, but other ions also condition the process. There is an optimum concentration for each salt; thus NaCl and KCl provoke polymerization at $0.1M$ concentration, but at high concentrations the process is reversed. Mg ion and ATP are necessary factors for G-F transformation, which does not take place in their absence. Polymerization has been followed with the electron microscope. Small globules of molecular dimensions first unite into larger ovoids 100 A in width and 300 A long; these are then joined end to end and form long filaments with a periodic cross striation at 300-A intervals.

[1] Engelhardt, W. A., and M. N. Ljubimowa, *Nature, London*, **144**, 669, 1939.

X-ray diffraction diagrams of muscles are the sum of the patterns given by extracted F-actin and myosin separately, and the two are oriented parallel to each other and to the muscle-fiber axis.[1]

Actomyosin. Myosin and F-actin in appropriate conditions of ionic equilibrium unite to form an F-actomyosin-ATP complex, a colloid of high viscosity. This complex contains 1 part of actin to 2.5 to 3 parts of myosin. The SH groups of myosin must be free for it to unite with F-actin.

If a thread of F-actomyosin is suspended in $0.05M$ KCl, and ATP (0.015 per cent) is added, the thread shortens and narrows very rapidly. Shrinking (syneresis of colloids) is due to loss of hydration water from the particles; in this case it takes place so quickly that it simulates contraction, except that the thread does not become shorter and wider like the muscle fiber, but simply shrinks. If, however, the actomyosin thread is stretched so that its micellae tend to be oriented along its axis, when ATP is added it does not shrink but becomes shorter and wider. The reaction is reversible, and the thread swells and stretches when the ionic strength is increased and ATP is added. Szent-Györgyi and his associates have shown that muscle fibers (rabbit's psoas) extracted with 50 per cent glycerol at 0°C. can be kept at low temperatures almost indefinitely without losing their contractility. If these fibers are submerged in muscle juice, or in a solution of KCl, Mg Cl_2, and ATP at the concentrations in which these substances are found in muscle juice, they contract. K^+ can be replaced by Na^+, but Mg^{++} cannot be replaced by Ca^{++}, and ATP is entirely specific; no other substance produces its effect. An increase in the concentration of KCl, Mg, ATP or a shift in pH will provoke dissociation of F-actomyosin into its components, and the fiber relaxes. A shift in concentration of $0.02M$ KCl is sufficient to dissociate or unite the two proteins.

The state of actin in the resting muscle is not known. Straub believes it is in the form of G-actin, which on excitation is converted to F-actin. Szent-Györgyi believes it is in the form of F-actin, but dissociated from myosin. The electron microscope shows filaments in the myofibrils running through the whole length of the sarcomere from one Z membrane to another.

[1] Astbury, W. T., *Nature, London*, **160**, 388, 1947.

These have been interpreted as made up of F-actomyosin or F-actin and myosin lying side by side. The A segment, however, contains more substance than the I segment; therefore Szent-Györgyi considers that the filaments are probably made up of F-actin, and that myosin is located only in the A segment which is the contractile part of the sarcomere. If this were so, "only half of the actin filament could take part in contraction, while the other half would merely act, so to speak, as a tendon for the contractile part,"[1] *i.e.*, would form part of the series elastic component.

Szent-Györgyi has suggested the following sequence of events to explain contraction and relaxation: The wave of excitation disturbs the ionic equilibrium so that conditions favorable to the union of actin and myosin ATP are created. F-actomyosin ATP in its extended state is unstable and tends to dissipate its thermodynamic potential, going into an energy-poorer shorter form. Dephosphorylation of ATP to ADP would bring about dissociation of actin and myosin and relaxation. During the period of recovery ATP is resynthesized and adsorbed by myosin. There is, however, no conclusive evidence as to whether hydrolysis of ATP is associated with contraction or with relaxation. The adequacy of this theory has been subject to criticism from other points of view, *e.g.*, the theoretically derived energetics of the system. "There still remains the bare mechano-chemical fact that actomyosin does somehow contract when acted on by ATP in the presence of physiological concentrations not only of ATP but of the cations K and Mg as well."[2]

Other proteins in muscle are the X globulin; tropomyosin, which is a prototype of myosin, perhaps one of the units from which the myosin filament is made; para-myosin; contractin appearing in contracting muscle; Y protein, which becomes strongly bound to the muscle stroma in contracted muscle; and N protein, a lipid-protein-nucleic acid complex showing negative birefringence, thus having the properties of the substance in the I band which extinguishes the positive double refraction of F-actin or actomyosin in this band.

Myohemoglobin (MHb) is found in higher concentration in red fibers (700 to 800 mg. per cent) than in white fibers; it increases after birth and decreases a little in senility. Its molecular weight is approximately 16,800, *i.e.*, that of one of the four units which make up hemoglobin. Evidence obtained from x-ray diffraction diagrams suggests it is a flat molecule made up of four parallel chains 54 A long, situated in the same plane; the width of the molecule is 37 A and its thickness between 9 and 14 A. This structure is similar to that of one of the four layers of hemoglobin. The affinity of myohemoglobin for oxygen is greater than that of hemoglobin and less than that of the cytochrome–cytochrome oxidase system; at 40 mm. Hg oxygen partial pressure, hemoglobin is only 38 per cent saturated, while myohemoglobin is 60 per cent saturated. The relative affinities for oxygen are such that myohemoglobin acts as an oxygen carrier, taking O_2 from the blood and giving it up to the muscle enzyme system.

Glycogen is found mainly in the anisotropous substance in close association with myosin. The total amount of glycogen in a resting muscle is approximately 500 mg. per cent; red fibers have two-thirds to three-fifths the glycogen content of white fibers. Glycogen diminishes in the course of contraction and falls to about 100 mg. per cent in a fatigued muscle; when the glycogen store has been exhausted, the muscle goes into contracture. Glycogen is rapidly restored during the period of recovery after exercise (see Chap. 41, Carbohydrate Metabolism). Hexoses, trioses, and several phosphoric esters of these sugars, and lactic and pyruvic acids are steps in the breaking down of glycogen, a process that sets free energy employed in muscular activity.

Phosphocreatine, also called creatinephosphate or phosphagen, acts as a phosphate donor on being split into creatine and phosphate. Creatinephosphoric acid is a stronger acid than phosphoric acid, therefore alkali is set free by this breakdown. More important still, a great amount of energy is liberated, from 10,000 to 12,000 cal. per mol. This is an "energy-rich" phosphate bond, to use a term employed by Lipmann, in contrast with others, such as hexosephosphate, which are "energy-poor," as on their being split only 3,000 cal. per mol is set free.

Adenosinetriphosphate or adenylpyrophosphoric acid (ATP) is also an important phosphate donor and acts as a coenzyme in phosphorylation reactions. The pyrophosphate is the active "energy-rich" bond. On being split, it gives one

[1] Szent-Györgyi, A., "Chemistry of Muscular Contraction," 2d ed., Academic Press, New York, 1951.
[2] Sandow, A., *Ann. Rev. Physiol.*, **11**, 297, 1949.

phosphate and adenosinediphosphate (ADP); in a second stage, another phosphate is set free and *adenosinemonophosphate* (AMP) or adenylic acid is formed (see "Intermediary carbohydrate metabolism," Chap. 41).

Diphosphopyridine-nucleotide functions as a hydrogen acceptor or coenzyme (cozymase I) to dehydrogenases acting on trioses. *Triphosphopyridine-nucleotide* (cozymase II) acts as a hydrogen acceptor in other reactions, *e.g.*, the oxidation of hexosemonophosphate to phosphohexonic acid (see Chap. 34).

Approximately fifty *enzymatic activities* have been found in muscle extracts (see Chap. 34), but the existence of an enzyme in minced muscle does not necessarily mean that it plays a part in the living fiber; it only suggests that it may do so. Moreover, it has been proved that certain chemical reactions that take place in minced muscle do not occur in the intact fiber.

The following are the main enzyme systems found in muscle:

1. *Cytochrome–cytochrome oxidase*, found in higher concentration in white than in red fibers (see Chap. 34).
2. *Flavoproteins* (Warburg's yellow respiratory pigment), the prosthetic group of which is flavindinucleotide (see Chap. 34 and "Vitamin B_2," Chap. 49).
3. *Dehydrogenases:* for lactate, hydroxybutyrate, succinate, malate, isocitrate, tartrate, glycerophosphate, glycero-aldehyde phosphate, triose phosphate, and glutamate.
4. *Adenosinetribhosphatases.* The ATP-ase activity of myosin inhibited by Mg^{++} with an optimum at pH 9, and of Mg actomyosinate with an optimum at pH 7, have already been mentioned. Another adenosinetriphosphatase has been extracted from muscle. It is independent of myosin and is activated by Mg^{++} but not by Ca^{++}.[1]
5. *Phosphorylase*, which activates the reversible hydrolysis (phosphorolysis) of glucose-1-phosphate (Cori's ester) in the synthesis and breakdown of glycogen (see Chap. 41). *Isophosphorylase* catalyzes the 1:6 linkages in the formation of glycogen.
6. *Phosphopherases, i.e.*, enzymes that catalyze transfer of phosphate from one compound to another (Parnas), such as *hexokinases* which transfer phosphate from ATP to glucose giving glucose-6-phosphate (Robison's ester) and to fructose giving fructose-6-phosphate (Neuberg's ester).[1] Another enzyme analogous to but different from hexokinase, activated by Mg^{++} and Mn^{++}, catalyzes the following reaction: glucose-1-phosphate + ATP ⇄ glucose-1:6-diphosphate + ADP.[2] Phospherases also transfer phosphate from creatinephosphate to ADP and AMP, and from ATP to Neuberg's ester giving fructose-1:6-diphosphate (Harden and Young's ester).
7. *Myokinase*, which dismutates ADP, activating the following reaction: 2 ADP ⇄ ATP + AMP.
8. *Phosphoglucomutase*, which catalyzes intramolecular transfer of phosphate from C^1 to C^6. Glucose-1:6-diphosphate, formed by the enzyme reaction mentioned above, acts as the prosthetic group of the enzyme. P is taken up by the enzyme from C^1 and returned to C^6 of the glucose-1-phosphate substrate,[3] as has been shown by means of radioactive P.[4] The reaction can be formulated as follows: glucose-1-phosphate + glucose-1:6-diphosphate ⇄ glucose-6-phosphate + glucose-1:6-diphosphate.
9. *Phosphohexose isomerase*, which catalyzes the equilibrium found in resting muscle between Robison's ester (70 per cent) and Neuberg's ester (30 per cent). The mixture of the two esters is known as Embden's ester.
10. *Aldolase*, which catalyzes the reversible cleavage of fructose-1:6-diphosphate into glyceraldehyde-3-phosphate and dihydroxyacetone phosphate.

Metabolic processes. When there is an adequate oxygen partial pressure in the muscle (aerobiosis), oxidation is ultimately the main source of energy. When the muscle contracts internal pressure is developed which is sufficiently high (100 to 300 mm. Hg) to suppress the blood flow through the capillaries. There is

[1] KIELLEY, W. W., and O. MEYERHOFF, *J. Biol. Chem.*, **176**, 591, 1948.

[1] Two hexokinases have been obtained from muscle and liver extracts; one is a fructokinase which acts on fructose but not on glucose, the other is a glucokinase (SLEIN, M. W., G. T. CORI, and C. T. CORI, *J. Biol. Chem.*, **186**, 763, 1950).

[2] PALADINI, A. C., R. CAPUTTO, L. F. LELOIR, R. E. TRUCCO, and C. E. CARDINI, *Arch. Biochem.*, **23**, 55, 1949.

[3] CARDINI, C. E., A. C. PALADINI, R. CAPUTTO, L. F. LELOIR, and R. E. TRUCCO, *Arch. Biochem.*, **22**, 87, 1949.

[4] JAGANNATHAN, V., and J. M. LUCK, *J. Biol. Chem.*, **179**, 560, 1949.

$HN{=}C(-NH-PO(OH)_2)-N(CH_3)-CH_2{\cdot}COOH$

Phosphocreatine

Adenine — Ribose — Triphosphate

Adenosinetriphosphate (ATP)

Nicotinic amide — Ribose — Diphosphate — Ribose — Adenine

Diphosphopyridine-nucleotide (DPN) (oxidized) (cozymase I)

Flavine — Ribose — Diphosphate — Ribose — Adenine

Riboflavin-nucleotide (oxidized)

satisfactory evidence not only that blood flow decreases during strong sustained contraction of human muscles,[1] but also that mechanical hindrance produced by rhythmic contractions reduces blood flow.[2] A condition of relative oxygen insufficiency (anaerobiosis) is set up, and phosphorolysis seems to be the main source of energy. Immediately after contraction has ended the blood flow increases considerably and the oxygen supply is greater than in the resting condition.

Aerobic processes. Respiratory metabolism. Muscle oxygen consumption increases considerably on contraction, and the increase is proportional to the activity developed. All those factors which condition development of tension in contraction also have an influence on oxygen consumption; *e.g.*, the length or tension of the resting muscle (initial length) conditions the resting oxygen consumption and the excess oxygen consumption due to activity.

The excess oxygen consumption of activity begins early in contraction. The first blood that can be obtained from a muscle after contraction has begun shows that there is an increased reduction of hemoglobin; therefore the muscle takes up more O_2. Millikan[3] has used a photometric method to follow the changes in oxygen saturation of myohemoglobin in the muscle fiber of the cat's soleus contracting *in situ*. Myohemoglobin begins to lose oxygen within the minimum time (200 msec.) in which the decrease can be registered after the commencement of contraction; within the first second this process has reached its maximum velocity. Desaturation of myohemoglobin probably starts even earlier. The soleus has a contraction time of 100 msec.; therefore if the increase in oxygen consumption begins during the phase of contraction, it must do so in a much shorter time than that recorded by Millikan. D. K. Hill[4] studied the oxygen consumption of the frog sartorius at 0°C., in order to decrease the velocity of the reactions; nearly all the excess O_2 consumption in this condition took place after the contraction phase had ended. The excess oxygen consumption persisting after activity has ceased has been called the "oxygen debt" by A. V. Hill. Recovery of the initial resting condition requires the energy set free by oxidation, and although the oxygen consumption diminishes rapidly after activity has ended, it remains for a long time above the resting level.

[1] Grant, R. T., *Clin. Sc.*, **3**, 157, 1938.

[2] Barcroft, H., and A. C. Dornhorst, *J. Physiol.*, **109**, 402, 1949.

[3] Millikan, G. A., *Proc. Roy. Soc., London, s.B.*, **123**, 218, 1937.

[4] Hill, D. K., *J. Physiol.*, **98**, 207, 1940.

Stannard[1] has observed that sodium azide and other inhibitors of the cytochrome system suppress oxidations due to activity without altering the resting oxygen consumption. Therefore there is more than one enzyme system that takes part in the oxidation processes of muscle. Several other facts are in agreement with this. Thus if the oxygen partial pressure is diminished so that it falls below 50 mm. Hg in venous blood (hemoglobin 85 per cent saturated), the oxygen consumption of the resting muscle in the dog decreases by one-third; a further reduction in O_2 partial pressure, down to 20 mm. Hg, has no effect. Therefore there is an enzymatic system that needs an O_2 partial pressure of more than 50 mm. Hg, while another remains active at much lower O_2 pressures.

Anaerobic processes. Muscle can contract in an atmosphere free of oxygen (anaerobiosis). The ratio of tension developed to initial heat is the same as in aerobiosis. The capacity to perform work is nevertheless considerably diminished. In aerobiosis the frog sartorius can contract at a rate of 23 per minute without showing signs of fatigue for more than 10,000 contractions; in anaerobiosis (the muscle being kept in oxygen-free Ringer solution) this same rhythm of contraction soon produces fatigue; a steady state can be kept up for a long time only if the frequency of contraction is not more than 5 or 6 per minute.[2]

In the course of his studies on fermentation, Pasteur[3] saw that yeast, when it cannot freely oxidize the sugar it uses for its nutritive processes owing to lack of oxygen, continues to live by obtaining energy from other exothermic chemical reactions, such as the formation of alcohol from sugar. This change to anaerobic fermentation when there is a deficiency of oxygen is known as the "Pasteur reaction." In muscle kept in anaerobiosis, glycolysis occurs, glycogen diminishes, and lactic acid is formed (it would be more correct to say lactate, as the lactic acid

[1] Stannard, J. N., *Am. J. Physiol.*, **126**, 196, 1939; **135**, 238, 1941.

[2] Hill, A. V., and P. Kupalov, *Proc. Roy. Soc., London, s.B.*, **125B**, 313, 1929.

[3] Pasteur, L., Etudes sur la bière, in "Oeuvres de Pasteur," vol. 5, Chap. VI, Paris, 1927.

is almost completely dissociated); lactic fermentation, therefore, takes place.

Fletcher and Hopkins's[1] classic work showed that in a resting muscle in aerobiosis there is very little lactic acid (0.015 gm. per cent). In anaerobiosis lactic acid increases gradually to 0.2 or 0.3 per cent, a process that can be speeded up and increased to 0.3 or 0.5 gm. per cent, by making the muscle contract by any means whatever—simple contraction, tetanus, heat, chloroform, etc. In the course of recovery, in an atmosphere of oxygen, lactic acid accumulated in anaerobiosis diminishes considerably (down to 0.07 gm. per cent). Meyerhof[2] studied the variations in muscle glycogen and lactic acid in relation to oxygen consumption. In aerobiosis glycogen diminishes without lactate being produced, and in both the resting and the active muscle the oxygen consumed corresponds to the amount needed to oxidize the glycogen that has disappeared. In anaerobiosis lactic acid is formed in quantities corresponding to the glycogen decrease. If the tissue is placed in oxygen after a period of anaerobiosis, lactic acid disappears and glycogen is resynthesized. Oxygen consumption in these conditions is greater than before the period of anaerobiosis, but the excess oxygen consumption (oxygen debt) would be sufficient to oxidize only one-fifth to one-fourth of the lactic acid that disappears, and glycogen increases in such quantities as if three-fourths to four-fifths of the lactic acid removed had been converted into glycogen. This series of chemical reactions, glycogen → lactic acid → glycogen, is known as the "Pasteur-Meyerhof cycle."

Several facts were discovered later that added considerably to the knowledge of the chemistry of muscular contraction:

1. In anaerobiosis a hexosemonophosphate accumulates, which was named "lactacidogen" by Embden, because it was thought to be the lactic acid precursor.
2. Fiske and Subbarow[3] and simultaneously the Eggletons[4] found a labile organic phosphorus compound, which disintegrates when the muscle contracts. This is phosphocreatine, called "phosphagen" by the Eggletons, because it is a source of the phosphate that appears on muscular contraction. During recovery in oxygen, phosphagen is totally resynthesized from creatine and phosphate. In anaerobiosis only 30 per cent of the phosphagen is resynthesized.
3. Lundsgaard[1] demonstrated that muscular contraction can take place without the formation of lactic acid in animals intoxicated with monoiodacetic acid. Iodacetate blocks especially the triosephosphate dehydrogenase. The intoxicated muscles contract for a short time, and when all the phosphocreatine has been disintegrated they enter into contracture. The drug does not modify the ratio of tension developed to heat production.
4. Myosin has adenosinetriphosphatase activity which splits off the terminal phosphate of ATP, thus breaking an energy-rich bond and releasing 11,000 cal./mol of phosphate. Myokinase catalyzes the reconstruction of this bond by the conversion of 2ADP into ATP and AMP. ATP is also resynthesized by the transfer of phosphate from creatinephosphate to AMP and ADP.

Chemical reactions that accompany muscular contraction. Glycogen breakdown is of primary importance in the process of liberating energy for muscular contraction. In the first stages phosphorolysis is the principal mechanism; oxidation takes place only in the final phases (see Intermediary Carbohydrate Metabolism, Chap. 41).

1. Glycogen is split by phosphorolysis into glucose-1-phosphate (Cori's ester). The reaction is activated by phosphorylase. Phosphorolysis does not entail a transfer of energy.
2. Glucose-1-phosphate is transformed to glucose-6-phosphate by intramolecular phosphate transfer; phosphoglucomutase with glucose-1:6-diphosphate catalyzes this reaction, as was explained in a previous paragraph.
3. Fructose-6-phosphate (Neuberg's ester) is formed by intramolecular rearrangement, so that an equilibrium of 70 per cent Robison's ester and 30 per cent Neuberg's ester is main-

[1] Fletcher, W. M., and F. G. Hopkins, *J. Physiol.*, **35**, 247, 1907.

[2] Meyerhof, O., *Pflüger's Arch. f. d. ges. Physiol.*, **176**, 88, 1919; **182**, 284, 1920; **185**, 11, 1920.

[3] Fiske, C. H., and V. Subbarow, *Science*, **65**, 401, 1927; *J. Biol. Chem.*, **81**, 629, 1929.

[4] Eggleton, P., and G. P. Eggleton, *Biochem. J.*, **21**, 190, 1927; *J. Physiol.*, **65**, 15, 1928.

[1] Lundsgaard, E., *Biochem. Ztschr.*, **217**, 162, 1930; **227**, 51, 1930.

tained in the resting muscle; phosphohexose isomerase catalyzes this equilibrium.

4. A further phosphorylation transforms fructose-6-phosphate to fructose-1:6-diphosphate (Harden and Young's ester); ATP again acts as a phosphate donor, with Mg^{++} activating the reaction.
5. Hexosediphosphate splits into two phosphorylated three-carbon-atom groups, probably dihydroxyacetone phosphate and 3-phosphoglyceraldehyde; aldolase catalyzes this reaction.
6. 3-Phosphoglyceraldehyde is phosphorylated and then oxidized to 1:3-diphosphoglyceric acid; a dehydrogenase activates this reaction and diphosphopyridine-nucleotide acts as hydrogen acceptor. The 1:3-diphosphoglyceric acid is dephosphorylated (ADP acting as phosphate acceptor) and converted by intramolecular phosphate transfer into 2-phosphoglyceric acid. Enolase activates the loss of H_2O, and phosphopyruvic acid results, which by dephosphorylation (ADP being the phosphate acceptor) gives pyruvic acid.

Glycogen
↓ *Phosphorylase*
Glucose-1-phosphate
↓ *Phosphoglucomutase*
Glucose-6-phosphate
↓ *Phosphohexose-isomerase*
Fructose-6-phosphate
ATP, Mg^{++} ↓ *Phosphohexokinase*
Fructose-1:6-diphosphate
↓ *Aldolase*
Dihydroxyacetonephosphate
+
3-Phosphoglyceraldehyde
DPN ↓ *Dehydrogenase* (oxidation)
1:3-Diphosphoglyceric acid
ADP ↓
3-Phosphoglyceric acid
↓ *Phosphoglyceromutase*
2-Phosphoglyceric acid
↓ *Enolase*
Phosphopyruvic acid
ADP ↓
Pyruvic acid
DPN ↓ *Dehydrogenase* (reduction)
Lactic acid

In a second phase pyruvic acid may follow one of two paths, according to the availability of oxygen. If there is oxygen deficiency (anaerobiosis), pyruvic acid accepts H_2 from reduced DPN and lactic acid is formed; this is a reversible reaction, catalyzed by a special dehydrogenase. In the organism lactic acid passes into the blood and in the liver is reconverted into glycogen, which by glycogenolysis is transformed into glucose, which is taken back to the muscle by the blood and there synthesized into muscle glycogen, which on being broken down again gives lactic acid. This series of reactions is known as the "Cori cycle." If, on the contrary, there is abundance of oxygen (aerobiosis), pyruvic acid is oxidized to CO_2 and H_2O, possibly along the path known as the tricarboxylic acid or citric acid cycle (see Chap. 34).

Creatinephosphate is disintegrated into creatine and phosphate, a highly exothermic reaction; the phosphate is taken up by ADP, and the reaction is catalyzed by a specific enzyme and Mg^{++}. This is a reversible process, and in the course of muscular contraction, there is first disintegration and then resynthesis of creatinephosphate. The resynthesis is an endothermic (endergonic) reaction; it therefore needs outside energy, which is obtained from glycolysis or oxidation. In monoiodacetic-acid intoxication the inhibition of these exothermic reactions prevents the reconstruction of phosphocreatine.

It is not yet possible to state exactly what chemical reactions take place in the living muscle fiber, nor in what order they come; neither can we correlate precisely chemical reactions and energy changes. From the facts known it seems that some process approximating the following sequence takes place in the living muscle: With the onset of contraction, there is a sudden and considerable increase in the need of oxygen, which causes a relative deficiency of oxygen and therefore a condition of anaerobiosis; this produces a "Pasteur reaction," *i.e.*, anaerobic fermentation. Anaerobic chemical reactions have been demonstrated in the course of the first minute of contraction in mammalian muscle with normal circulation: lactic acid, hexosemonophosphate, and more slowly inorganic phosphate increase, while glycogen and creatinephosphate diminish rapidly and adenosinetriphosphate more slowly.[1] Oxygen supply soon improves owing to vasodilatation and the opening of capillaries that are closed in the

[1] FLOCK, E. W., D. S. SINGLE, and J. L. BOLLMAN, *J. Biol. Chem.*, **129**, 99, 1939.

resting muscle. A steady state is then established, during which lactic acid concentration in muscle returns to the resting value, owing principally to diffusion into the blood. Hexosemonophosphate and more slowly ATP and inorganic phosphate also return to the resting values. Glycogen and phosphocreatine remain low. This steady state can be kept up for a long time. When glycogen is exhausted, the muscle enters into a lasting contracture (fatigue cramp). Oxidation reactions are the principal source of energy; anaerobic processes are probably emergency reactions set in motion when there is oxygen deficiency, *i.e.*, at the beginning of contraction.

Changes in acid-base equilibrium during contraction. Chemical reactions in muscle produce rapid shifts in the acid-base equilibrium. In order to follow them, their velocity can be diminished by cooling. In frog muscle at 0°C., as soon as contraction begins there is a shift toward the alkaline side. Then there is a shift in the opposite direction, attributed to disintegration of ATP. A progressive alkalization follows, due possibly to phosphocreatine disintegration; this takes place mostly when the muscle is already relaxed. Finally there is a prolonged acidification, attributed to lactic acid formation, because it is more marked in anaerobiosis and does not take place in muscles intoxicated with monoiodacetate.[1]

Owing to the delay with which these shifts are recorded, it is impossible to correlate them with the mechanical. chemical, or thermic phenomena of contraction. "No present method is nearly rapid enough, nor probably sensitive enough to record changes of pH suspected to occur in a single contraction."[2]

ENERGETICS OF MUSCULAR CONTRACTION

In the resting muscle, as in all living tissues, there is a continuous evolution of energy. Activity causes a considerable increase in the mobilization of energy, because muscle is a machine which converts chemical energy into mechanical energy. Energy evolved can be measured as heat, but the amount of heat produced in a muscle twitch is very small, about 3 μcal per gm. Heat production of muscle is a complicated phenomenon; to analyze it, it is usually necessary to record it as a function of time, but a twitch, even in a cooled muscle, is a very rapid event; *e.g.*, in frog's skeletal muscle at 0°C. heat begins to appear about 10 msec. after the stimulus has been applied, reaches its peak in 30 to 40 msec., and is complete in about 400 msec. The recording instruments should, therefore, be highly sensitive and very rapid. Hill[1] has developed a thermopile of palladium-gold and iron with a very rapid galvanometer which can register with a sensitivity of 10^{-5}°C. and a time lag as small as 1.5 to 4 msec. The mechanical events (shortening or tension) can be recorded simultaneously.

Heat is produced in the course of muscle activity in two main phases: (*a*) during the mechanical response, *initial heat;* (*b*) after mechanical response, *recovery heat.*

Initial heat.[2] Heat produced during the mechanical response in a single twitch is composed of two parts: (*a*) heat of activation, and (*b*) heat of shortening; the observed heat is the sum of the two. If the stimulus is sustained, *i.e.*, in a tetanus, a third component appears, (*c*) heat of maintenance.

Heat of activation is calculated by subtracting from the observed heat the heat of shortening. It is rather less than the maximum heat of shortening and is not dependent on the length at which the muscle is stimulated or changes in length thereafter, or on the load, or on work done. It begins during the latent period of the mechanical response well before shortening or the rise in tension commences; thus the latent period of activation heat is about 10 msec. in the frog's muscle at 0°C., and the mechanical latency is 20 to 25 msec.; in the tortoise's muscle the respective figures are 60 msec. and 90 to 100 msec. The heat of activation starts at a maximum rate, then diminishes in rate continuously as contraction proceeds.

Heat of shortening begins when the muscle commences to shorten and is proportional to the amount of shortening. If the muscle does not shorten, the heat observed is equal to the heat of activation. Heat of shortening is not dependent on the load, speed of shortening, or work done.

The total energy liberated in a muscle twitch can be expressed as $E = A + W + ax$, where A is the activation heat, W is the work, and ax is the heat of shortening. This relation is true for the whole contraction and for any part of it.

[1] HILL, D. K., *J. Physiol.*, **98**, 467, 1940. DUBUISSON, M., *Proc. Roy. Soc., London, s.B.*, **137**, 67, 1950.
[2] HILL, A. V., *Proc. Roy. Soc., London, s.B.*, **137**, 40, 1950.

[1] *Ibid.*, **136**, 228, 1949.
[2] *Ibid.*, **136**, 195, 220, and 242, 1950; **137**, 268 and 330, 1951.

The rate at which energy is liberated, in excess of activation heat, in a muscle twitch and in a maintained tetanic contraction is a decreasing linear function of the load. This should be connected with the characteristic relation between force and velocity discussed above (see "Shortening," under "Mechanical phenomena of muscular contraction").

Heat of maintenance in a tetanic contraction is the sum of the heat of activation produced by each one of the successive shocks of the stimulus.

Heat produced during relaxation is equivalent to the mechanical energy which disappears at the same time. If the muscle is without load or tension no heat is produced after shortening is ended. The process of relaxation itself, therefore, is not accompanied by the production of heat.[1]

Recovery heat. Heat continues to be produced after the mechanical response is ended at a low rate and for a long time. The absence of oxygen does not modify the output of initial heat, but diminishes considerably the heat of recovery. In the presence of oxygen total recovery heat is about equal to total initial energy as heat and work together.

Contraction may be regarded as due to entropy changes as in the shortening of rubber, or to changes in potential energy resulting from chemical reactions. In an entropy cycle the strength of intermolecular attractions should decrease during contraction, while in a potential energy cycle they ought to increase; observations in muscles suggest that they do decrease.[2] There are, however, fundamental objections to the entropy theory. If shortening is due to muscle passing from a state of low entropy to one of high entropy, as in the retraction of stretched rubber, the heat of shortening should be negative. This is not so, since heat is given out during shortening; therefore, from the thermodynamic point of view, entropy is decreased by contraction. When rubber shortens, if it does work, the heat absorbed is greater in proportion to the work done, while in muscle heat of shortening is not affected by work.[3]

Efficiency. Mechanical efficiency is the ratio of the energy equivalent of work performed to the total energy mobilized (see Chap. 47). The gross efficiency of the whole organism is 17 to 20 per cent, *i.e.*, less than one-fifth of the total energy spent can be recovered as work; the rest is given off as heat. The maximum efficiency of a muscle twitch or a tetanic contraction is approximately 40 per cent, according to results obtained with accurate myothermic methods.[1] Previous determinations had given lower figures.

[1] *Ibid.*, **136**, 211 and 420, 1950.

[2] Pryor, M. G. M., *Proc. Roy. Soc., London, s.B.*, **137**, 71, 1950.

[3] Hill, A. V., *Proc. Roy. Soc., London, s.B.*, **137**, 49, 1950.

SMOOTH MUSCLE

Smooth muscles are made up of elongated, spindle-shaped cells, with a central nucleus and a thin membrane. The fibers are much shorter than those of striated muscle. Each one has its origin in a single mesenchyme cell, or in an ectodermal cell (piloerectors, iris). They are not multinucleated giant cells as are the striated muscle fibers. The protoplasm is differentiated into slender fibrils which have no transverse striation, but in the adductor muscle of the clam regular segments repeated periodically at intervals of 1100 A have been observed by means of the electron microscope. In some smooth muscles, the fibrils pass from one cell to another; the protoplasmic bridges endow these muscles with some of the characteristics of a syncytium. Myosin, which differs slightly from striated-muscle myosin, is the principal constituent of the fibrils and gives the cell the property of birefringence, its anisotropy being similar to that of the A or Q disk. Microscopic examination of living smooth-muscle cells, in tissue cultures, has shown a displacement on contraction of the peripheral protoplasm (Roskin's kinoplasm) toward the perinuclear region. The middle part of the fiber is thickened and its volume diminishes; therefore the interfibrillar space increases. The protoplasm loses its transparency, becomes cloudy, and stains more deeply, a sign of condensation.

Smooth muscle forms membranous structures, sheets, or ribbons, such as the nictitating membrane, the retractor penis, and the piloerectors; in other cases it forms part of the walls of cavities or ducts, such as the bladder, the uterus, the digestive tract, the ureter, the blood vessels, etc. Smooth muscles do not constitute a homogeneous group; they have the property of contractility in common, but excitation and conduction vary considerably from one muscle to another. Bozler[2] distinguishes two types of smooth muscle:

[1] *Ibid.*, **136**, 220, 1950.

[2] Bozler, E., *Biol. Symposia*, **3**, 95, 1941.

1. *Multiunit smooth muscles*, which are similar to striated muscle. They are made up of motor units and are usually activated by motor nerves. The nictitating membrane, the piloerectors, and the muscles of the blood vessels belong to this group.
2. *Visceral smooth muscles*, which are similar to cardiac muscle. They are syncytial in character and have a well-developed automatism, being less dependent on the nervous system than the previous group. Among these are the muscles of the intestinal tract and the uterus.

Innervation. Smooth muscles are innervated by sympathetic and parasympathetic postganglionic fibers. These are fine, nonmyelinated fibers, with the exceptions of those from the ciliary ganglion which have a myelin sheath up to their ending in the intrinsic musculature of the eye. The nictitating membrane and the piloerectors are innervated exclusively by the sympathetic and the ciliary muscle by the parasympathetic; in these muscles the nerve impulses produce contraction. Most smooth muscles have a double innervation; one of the nerves has a stimulatory, and the other an inhibitory, effect. These smooth muscles resemble the muscles of invertebrates, which also have excitatory and inhibitory innervation. They differ from striated muscles, which receive only excitatory impulses, inhibition being due to the suppression of the central excitatory state (central inhibition).

Each fiber of the ciliary muscle has a nerve ending, but in the majority of smooth muscles not all the fibers are innervated.[1] The cells that are not innervated are stimulated indirectly, perhaps by the chemical mediator, sympathin or acetylcholine. The motor units in the multiunit muscles are not so well defined as those in striated muscle. Stimulation of an axon going to a striated muscle provokes contraction in a definite number of muscle fibers, namely, all those innervated by the peripheral branching of the axon; section of the nerve fiber suppresses the activity of these same muscle fibers. Consequently there is a constant ratio between the responses to stimulation of a fixed fraction of the nerve and of the whole nerve, whatever the frequency of stimulation. In the nictitating membrane, on the contrary, as frequency of stimulation increases, a larger number of muscle fibers respond.[1] This may be due to the liberation of a greater amount of chemical mediator (sympathin) by the more frequent stimulus and therefore the activation of a larger number of muscle fibers.

[1] CANNON, W. B., and A. ROSENBLUETH, "Autonomic Neuro-effector Systems," Macmillan, New York, 1937.

Nerves have a twofold action on the smooth muscles: (*a*) nerve impulses liberate chemical mediators and thus have a stimulatory or inhibitory effect; (*b*) nerves also have a "trophic" action, which is made manifest by denervation, *e.g.*, nerve degeneration sensitizes the nictitating membrane, and other adrenergic effectors, to adrenaline. This effect is due to the suppression of a direct action of the nerve on the muscle fiber, because when only a few fibers of the nerve going to the nictitating membrane are cut, only the units that have lost their innervation are sensitized.

The effects of several chemical substances are of great physiological, as well as pharmacological or therapeutic, interest (see Chap. 84). They can be classified into three groups: (*a*) sympathicotropic; (*b*) parasympathicotropic; (*c*) musculotropic. The first two act on the receptors for the sympathetic and parasympathetic chemical mediators and produce the same effects as these. The third group of substances acts directly on the muscle fiber; among these is the extract of the posterior hypophyseal lobe.

Excitation and conduction. The excitability of smooth muscle is remarkable for (*a*) its slowness—chronaxies are very long, *e.g.*, 3 to 5 msec. for the stomach of the frog and 1.5 sec. for the blood vessels in the web of the frog; (*b*) its instability—many factors can change it.

From the standpoint of excitation and conduction, smooth muscles are not a homogeneous group of effectors. Certain muscles, *e.g.*, the nictitating membrane, the piloerectors, and the nonpregnant uterus of the cat in anestrus give little or no response to direct electric stimulation. Stimulation of the nerve provokes a response after a long latent period. Neuromuscular delay is about 30 msec. in the nictitating membrane and 40 msec. in the piloerectors; mechanical latency is approximately 150 msec. A single shock may be followed by a single electrical and mechanical response, similar to a slow twitch, lasting ten to fifteen times as long as the twitch of the soleus, which is a "slow" striated muscle.

[1] ROSENBLUETH, A., and D. M. RIOCH, *Am. J. Physiol.*, **106**, 365, 1933.

More frequently a rhythmic response is observed; repeated contractions with the corresponding electric potentials at a rate of one or two per second produce an effect resembling an incomplete tetanus. Adrenaline provokes a similar response. A propagated excitatory wave,

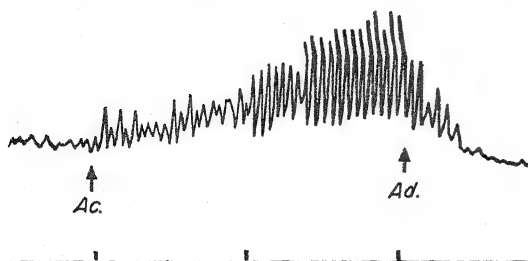



FIG. 352. Rhythmic contractions of the isolated duodenum of the rabbit. *Ac.*, acetylcholine 4×10^{-6}; *Ad.*, adrenaline 5×10^{-6}. Lower line, time in minutes.

with a spike potential, followed by a retractory period has not been satisfactorily demonstrated in this type of muscle; neither do they follow the all-or-nothing law. The nerve impulse does not stimulate the muscle fibers directly—their electrical excitability is far too low for that—but releases a chemical mediator. This mediator acts directly on the contractile element of the muscle fiber; it does not evoke a propagated disturbance such as is observed in striated muscle, which spreads throughout the fiber and provokes contraction. Contraction is graded in relation to the strength of the stimulus, *i.e.*, to the amount of mediator which is released; it is not therefore an all-or-nothing response.

Other smooth muscles respond to direct electric stimulation, *e.g.*, intestinal muscle, the ureter, the cat uterus during estrus, and the rabbit uterus. The response to local stimulation is propagated throughout the muscle accompanied by a spike potential; it is followed by a refractory period and is an all-or-nothing phenomenon. In the uterus in estrus spontaneous rhythmic waves of contraction, which are very prominent, also spread throughout the uterine horn. Stimulation of the nerves to these muscles evokes contraction or relaxation, according to the nerve and muscle, and the condition of the latter. Thus the nonpregnant uterus of the cat not in estrus always responds to stimulation of the hypogastric nerve with relaxation. In the rabbit, stimulation of nerves distributed to limited regions of the uterus provokes contraction which does not spread throughout the muscle. The vagus nerve and its chemical mediator, acetylcholine, have a predominantly excitatory effect on the motility of the intestine, evoking or enhancing rhythmic contractions and propagated contractions (peristalsis). The sympathetic, and adrenaline, have a predominantly inhibitory effect[1] (Fig. 352). Electrical excitability, spontaneous activity and conduction are not dependent on ganglion cells or nerve fibers. They have been observed in muscles after the nerves have been cut and have degenerated, and in intestinal strips in which subsequent serial section showed there were no nerve cells. Moreover a 1 per cent solution of cocaine suppresses all nervous activity but does not modify conduction or rhythmic contractions.

Rosenblueth,[2] on the basis of transmission and activation, divided smooth muscles into two classes, which differ from those of Bozler: (*a*) muscles in the first class, *e.g.*, the nictitating membrane and the piloerectors, have short fibers, there is no conduction, and the chemical mediator acts directly on the contractile substance, without an intermediate propagated disturbance which spreads excitation throughout the muscle fiber; (*b*) muscles in the second class have long fibers, and the mediator evokes a propagated process which acts on the contractile substance.

Contraction of smooth muscle. Contraction of smooth muscle can be studied *in situ* by registering the movements of the viscera of which it forms part, or in isolated strips submerged in an adequate saline solution (*e.g.*, Ringer's or Locke's), in which they survive and respond for many hours. Contraction in response to stimulation of the nerve, an electric stimulus, or a chemical substance, is very slow and frequently rhythmical. The contraction time of a single twitch is 10 to 15 times that of slow striated muscles. Relaxation is even slower, but can be speeded up by an inhibitory stimulus. The maximum tension developed in contraction increases with the resting length up to an optimum, as in striated muscle; but this optimum initial length varies considerably and is influenced by the previous condition of the muscle.

[1] The effects of nerve stimulation and of the chemical mediators on the sphincters are opposite to those they have on the wall muscles.

[2] ROSENBLUETH, "The Transmission of Nerve Impulses at Neuroeffector Junctions and Peripheral Synapses."

Repeated stimulation provokes fatigue; contraction and relaxation are shortened, and finally there is complete relaxation instead of contracture such as occurs in fatigued striated muscles.

When a smooth muscle is stretched it lengthens, at first rapidly and then more slowly, taking a long time to reach the final length, owing to internal friction or "viscosity." Tension also increases on stretching, but after quickly arriving at a maximum, it decreases slowly following an exponential curve, similar in its time course to that of relaxation. This release of tension without change in length gives great "plasticity" to smooth muscles, a property of value in visceral functions. The distention of a cavity with walls made up of smooth muscle, by the increase of its contents, at first raises the internal pressure, and then, without any change in its capacity, the pressure falls because of the release of tension in the muscles. Sudden distention can produce a greater and more lasting increase in the internal pressure, because it acts as a stimulus and provokes contraction of the wall muscles.

Tonus. This is an outstanding feature of smooth muscle, but the term is used for several different conditions. Evans defines it by the ratio $F:L$, in which F is the stretching force and L the length of the muscle. It varies considerably from one moment to another and there is no constant ratio between tension and length, such as is observed in striated muscle. Tonus is originated in several different ways: (*a*) it can be due to asynchronous activity in different parts of the muscle; (*b*) it can be produced by nerve impulses of very low frequency (2 per second); (*c*) it can arise in the muscle fiber itself, a fact proved by the tonus observed in strips of muscle free from nerve cells. Tonus in smooth muscle is therefore not an exclusively reflex phenomenon, which can be suppressed by denervation, like the tonus of striated muscle.

Conditions that modify the activity of smooth muscle. Excitability, conduction, and contraction of smooth muscle are subject to considerable variations due to the influence of different factors. An increase in *temperature* diminishes excitability and tonus; between 45 and 50°C. relaxation is complete, the thermic contracture seen in striated muscle at a temperature of 45°C. is not observed. Cold increases tonus and prolongs the survival of smooth muscle with all its properties for several days. *Acidity* produces relaxation and quiescence in smooth muscle; in striated muscle it provokes contracture. *Alkalinity* increases tonus and spontaneous rhythmic activity. *Ionic equilibrium* plays an important part in the activity of smooth muscle. A decrease in the total concentration of salts increases excitability and the amplitude of rhythmic movements; an increase has the opposite effect. K^+ in adequate concentration increases tone and contractions; Ca^{++} has opposite and antagonistic effects to K^+. The whole response of smooth muscles is conditioned by the K:Ca ratio.

The *previous activity* of the muscle has considerable influence on the response to a subsequent stimulus. A contracted muscle or one that is in a condition of high tonus will relax when excited by a stimulus that in other conditions would provoke contraction. The condition of the muscle can also modify the effects of nerve stimulation. Thus stimulation of the vagus usually produces an increase in tonus and in the rhythmic movements of the intestine, but an inhibitory effect may be observed if, when the nerve is stimulated, there is already a high degree of activity. The sympathetic, which usually inhibits, can stimulate the relaxed intestine.

Hormones exercise considerable influence on the activity of smooth muscles. For example, the uterus of the cat in estrus has rhythmic contractions and responds to electric stimulation with a propagated contraction; during anestrus there is little or no activity and it is difficult to obtain a response to electric stimulation. The injection of estrogen in a cat in anestrus induces in the uterus the conditions peculiar to estrus. Estrone injections in women increase the tonus and diminish the amplitude of the spontaneous contractions of the uterus; progesterone has the opposite effect.

Chemical phenomena. Smooth muscle has a chemical structure similar to that of striated muscle. Myosin is the principal protein, but it is not the same myosin as that of striated fibers. Glycogen is found in a concentration of about 150 mg. per cent. Inorganic phosphate makes up half the total P; a large quantity of P is bound in nucleoproteins. Creatinephosphate is found in a quantity approximately one-tenth that of striated muscle. There is less potassium than in striated muscle; the Na:K ratio is 1:1.5 instead of 1:5. There is no stable basal condition in smooth muscle because tension, length, and rhythmic contractions are subject to

almost continuous variations. Oxygen consumption rises and falls with the activity of the muscle; there is, therefore, no strictly basal oxygen consumption. Oxidative reactions are necessary for the maintenance of tonus, and a smooth muscle relaxes if placed in a nitrogen atmosphere or if cyanide is added to the fluid in which it is submerged. Oxygen restores tonus which has been lost in anaerobiosis. If a smooth muscle is stretched, length, tonus (resistance to stretching), and oxygen consumption increase.[1] If it shortens, *e.g.*, in the course of isotonic contraction provoked by histamine, oxygen consumption diminishes,[2] but not if the muscle contracts against a resistance and tension is developed on shortening. Oxygen consumption varies in the same direction as tonus in the intestinal muscle maintained in isometric conditions. Stretching increases oxygen consumption to a greater or lesser degree, according to the resistance opposed to stretching and the changes caused in rhythmic movements. The transitory increase in tonus which can be provoked by histamine or acetylcholine is accompanied by an increase in oxygen consumption. The decrease in tonus caused by atropin, and in certain conditions by adrenaline, is accompanied by a decrease in oxygen consumption.[3] These facts show that in smooth muscle, at all events in those of Bozler's visceral type, there is a close relationship between tonus (resistance to stretch) and metabolic activity.

In anaerobiosis glycolysis occurs and lactic acid accumulates. After a period in anaerobiosis oxygen consumption rises above the initial level (payment of oxygen debt), lactic acid diminishes, and glycogen increases. Phosphocreatine (phosphoarginine in invertebrate muscle) diminishes in anaerobiosis and is resynthesized during a subsequent period of aerobiosis.

[1] Bayliss, L. E., *J. Physiol.*, **65**, 1P, 1928.
[2] Evans, C. L., *J. Physiol.*, **58**, 522, 1923.
[3] Bülbring, E., *J. Physiol.*, **122**, 111, 1953.

Smooth muscle has the same fundamental properties as striated and cardiac muscle, but its reactions have not the constancy and uniformity observed in other kinds of muscle, because it does not form a homogeneous group and because it is conditioned by a large number of external factors. Its importance resides in the part it plays in visceral functions.

BIBLIOGRAPHY

Bozler, E., Conduction, Automaticity, and Tonus of Visceral Muscles, *Experientia*, **4**, 213, 1948.

Buchtal, F., Muscle, *Ann. Rev. Physiol.*, **9**, 119, 1947.

Evans, C. L., The Physiology of Plain Muscle, *Physiol. Rev.*, **6**, 358, 1926.

Fenn, W. O., Muscle, *Ann. Rev. Physiol.*, **3**, 209, 1941.

Fischer, E., Muscle, *Ann. Rev. Physiol.*, **5**, 133, 1943.

Fischer, E., Vertebrate Smooth Muscle, *Physiol. Rev.*, **24**, 467, 1944.

Fulton, J. F., "Muscular Contraction and the Reflex Control of Movement," Williams & Wilkins, Baltimore, 1926.

Gasser, H. J., Contractures of Skeletal Muscle, *Physiol. Rev.*, **10**, 35, 1930.

Hill, A. V., *et al.*, A Discussion on Muscular Contraction and Relaxation; Their Physical and Chemical Basis, *Proc. Roy. Soc., London, s.B.*, **137**, 40, 1950.

Hill, A. V., *et al.*, A Discussion on the Thermodynamics of Elasticity in Biological Tissues, *Proc. Roy. Soc., London, s.B.*, **139**, 469, 1952.

Hill, D. K., Muscle, *Ann. Rev. Physiol.*, **13**, 99, 1951.

Millikan, G. A., Muscle Hemoglobin, *Physiol. Rev.*, **19**, 503, 1939.

Millikan, G. A., The Chemistry of Muscle, *Ann. Rev. Biochem.*, **11**, 497, 1942.

Muscle, *Biol. Symposia*, **3**, 1941.

Needham, D. M., The Biochemistry of Muscle, *Ann. Rev. Biochem.*, **6**, 395, 1937.

New York Academy of Sciences, **47**, 675–929, 1947 (symposium on muscular contraction).

Sacks, J., Changing Concepts of the Chemistry of Muscular Contraction, *Physiol. Rev.*, **21**, 217, 1941.

Sandow, A., Muscle, *Ann. Rev. Physiol.*, **11**, 297, 1949.

Soskin, S., and R. Levine, "Carbohydrate Metabolism," University of Chicago Press, Chicago, 1946.

Szent-Györgyi, A., "Chemistry of Muscular Contraction," 2d ed., Academic Press, New York, 1951.

CHAPTER 68

The Conduction of the Nerve Impulse

THE NERVOUS SYSTEM transmits the excitatory state from one part of the organism to another. One of the aspects of this function, transmission within the neuron, can be studied in the axons that make up a nerve. Transmission from cell to cell through the synapse will be studied later.

Conduction along the axon is of two different types: (*a*) the so-called trophic influence, which irradiates from the perikaryon and maintains the structural and functional integrity of the fiber; (*b*) the nerve impulse, or propagated excitatory disturbance. Very little is known about the nature of (*a*) (see Chap. 65). With respect to (*b*), two different properties are usually considered in a nerve fiber: excitability, or the capacity to respond to a stimulus, and conductivity, or the capacity to transmit the excitatory state. Two different processes can, in fact, be distinguished, the *local excitatory state* and the *propagated excitatory state*, but both are phases of the same phenomenon. Excitation and conduction are produced by the same fundamental processes in all the axons that have been examined, but there are important quantitative differences between different types of fiber even in the same species.

In physiologic conditions a neuron is stimulated exclusively by a receptor, another neuron, or a physicochemical change in the nerve center. The axons that form part of a nerve do not stimulate each other; they are insulated conductors, and this insulation would have to be suppressed for the excitatory state to spread laterally from one axon to another. The excitatory state is thus propagated in an orderly way; impulses from different receptors are transmitted to the appropriate centers, are recognized and localized, although neighboring axons may be transmitting impulses from other receptors to other centers. In like manner the motor units are stimulated separately and only by the central excitatory state.

The excitatory state of one fiber, however, has a certain influence on the excitability of neighboring fibers. This was first seen in unmyelinated fibers of invertebrates and was later observed in the sciatic nerve of the frog[1] and the peroneal nerve of the cat.[2] The action potential of the conducting fiber produces electrotonic currents in neighboring fibers and thus provokes alternating changes in the threshold. At first there is depression (anelectrotonus), followed by an increase of excitability (catelectrotonus) which coincides with the ascending phase of the spike potential of the conducting fiber, and finally a second period of depression (Fig. 353). These rapidly alternating states are sometimes followed by a more prolonged lowering of the threshold. Rosenblueth suggests that the early changes are due to electrical, and later ones to chemical, factors, among which potassium probably plays an important part. When two axons are stimulated simultaneously the speed of conduction is less than when one alone is stimulated, because the impulses progress into parts of the fibers in which the excitability is depressed by the anelectrotonus produced by the action potential of the neighboring fiber. When two fibers are stimulated successively, the speed of conduction decreases in the one stimulated first and increases in the other; thus there is a tend-

[1] MARRAZZI, A. S., and R. LORENTE DE NÓ, *J. Neurophysiol.*, 7, 83, 1944.

[2] ROSENBLUETH, A., *Am. J. Physiol.*, **140**, 656, 1944.

ency to synchronization of the impulses conducted in the several fibers of the nerve.[1]

The local excitatory state. A nerve is excitable in all its length. At a stimulated point a local excitatory condition arises, which has the following features: (*a*) the size depends on the strength of the stimulus, and does not respond to the "all-or-nothing" law; (*b*) it diminishes as the distance to the locus of stimulation increases, following an exponential curve; (*c*) it decreases gradually at a velocity defined by a constant that is directly proportional to the chronaxie; (*d*) it is not followed by an absolute refractory state; (*e*) the effect of a subsequent excitation can be added to it. When this local state has developed to a sufficient size and extended over a certain area, a propagated disturbance, the nerve impulse, is fired off and transmitted in both directions along the whole neuron (see Chap. 66).

Fig. 353. Interaction of nerve fibers. Changes in the excitability of a fiber (thin line) when an impulse is conducted by a neighboring fiber are superimposed on the spike potential of the conditioning fiber. (*Marrazzi, A. S., and R. Lorente de Nó, J. Neurophysiol., vol. 7, p. 83, 1944.*)

[1] Katz, B., and O. H. Schmitt, *J. Physiol.*, **97**, 471, 1940.

The nerve impulse. The propagated excitatory state is also called the nerve impulse. It is accompanied by metabolic phenomena such as an increase in respiratory exchanges and in heat production, the liberation of acetylcholine, changes in the permeability and the electrical properties of the membrane, and several electrical potentials. The latter have been the subject of a vast amount of work, which has yielded results of great value for the understanding of nerve physiology.

ELECTRICAL PHENOMENA OF THE NERVE IMPULSE

Electrical potentials that occur when a nerve is stimulated and conducts an impulse have been registered with great precision. Adrian and his associates have recorded electrical potentials in a single axon by dissecting a nerve and cutting it down until only one or a few fibers were left,[1] or by stimulating a receptor connected with a single afferent fiber.[2] Erlanger and Gasser[3] have made a remarkable contribution to nerve physiology by the use of the cathode-ray oscillograph. The potentials registered in a nerve are the sum of the individual axon potentials. Close to the point stimulated, all the fibers enter simultaneously into activity, but as the impulse is conducted along the nerve important changes occur, mainly because of differences in the conduction velocities of the fibers. The analysis of these potentials has given much valuable information on the properties of nerve fibers.

Local potential. When a nerve is stimulated by a brief electric shock, an electrotonic potential is produced which should not be mistaken for the shock artefact. This potential extends beyond the point where the electrode is in contact with the nerve. Its intensity is proportional to the strength of the stimulus. It diminishes as the distance to the electrode increases, following an exponential curve, and it decays, also exponen-

[1] Adrian, E. D., and D. W. Bronk, *J. Physiol.*, **66**, 81, 1928; **67**, 119, 1929.

[2] Matthews, B. H. C., *J. Physiol.*, **71**, 64, 1931; **72**, 153, 1931.

[3] Erlanger, J., and H. S. Gasser, "Electrical Signs of Nervous Activity," University of Pennsylvania Press, Philadelphia, 1937.

tially, with a time constant that is characteristic for each type of nerve fiber. The local potential is due to depolarization, and its intensity, spread, and decay correspond in space and time to the local excitatory state provoked by the same stimulus.

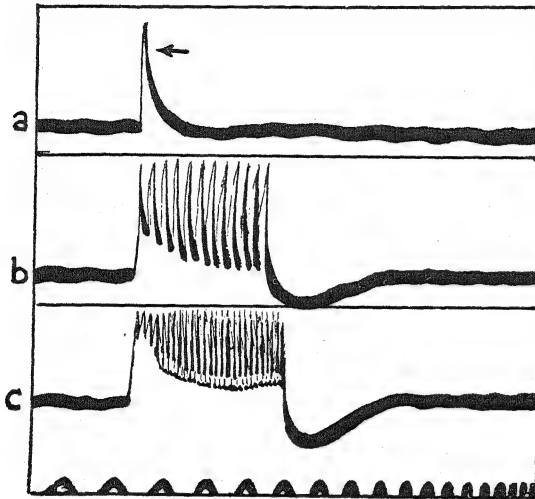



FIG. 354. Action potentials of phrenic nerve. *a*, single stimulus; *b* and *c*, repeated stimulation at rates of 180 and 350 per second. There is a marked positive after-potential in *b*, which is even more marked in *c*. Time, 20 msec. (*Erlanger, J., and H. S. Gasser, "Electrical Signs of Nervous Activity," University of Pennsylvania Press, Philadelphia, 1937.*)

"Spike" propagated potential. When the local potential reaches a critical level (the threshold) it increases suddenly, and a "spike" potential is transmitted along the nerve together with the propagated excitatory state. The electrotonic potential and the spike are independent of each other and vary in different ways under the influence of several factors. Thus cold does not change the electrotonic potential, but prolongs the duration of the spike potential. The spike potential is an "all-or-nothing" phenomenon, *i.e.*, its size is not dependent on the strength of the stimulus, but on the condition of the nerve. It is the best sign of a conducted impulse. The voltage is approximately 100 mv.; the ascending phase lasts 1.5 msec. in unmyelinated axons and 0.3 to 0.4 msec. in the most rapid myelinated axons of vertebrates. The descending phase is more prolonged, but it is not easy to mark exactly where it ends, because it overlaps other potentials, known as after-potentials (Fig. 354).

After-potentials. The spike is followed by a *negative after-potential.* It begins during the descending phase of the spike. Its amplitude is only 3 to 5 per cent that of the spike. It lasts approximately 15 msec. in the fastest fibers of mammals and 50 to 80 msec. in slower fibers.

A *positive after-potential* then follows. It is composed of two elements, sometimes called the "first positive" (P_1) and the "second positive" (P_2) after-potentials. The amplitude of this potential is only 0.2 per cent that of the spike in the fastest fibers and 1.5 to 4 per cent in slower fibers. Its duration varies from 40 to 60 msec. in fast fibers to more than 1,000 msec. in slow fibers.

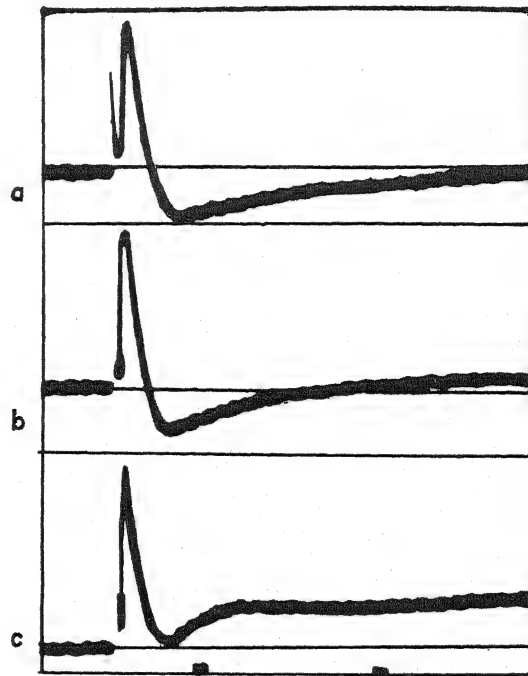



FIG. 355. Action potentials of a C fiber of a frog intoxicated with veratrine. The first spike, which is scarcely visible, corresponds to other fibers of higher conduction velocity. The well-marked spike corresponds to the C fibers. *a*, normal nerve, showing pronounced positive after-potential; *b*, veratrine increases the negative after-potential, which begins to mask the positive after-potential; *c*, at a later stage the negative after-potential is so great that it has completely masked the positive after-potential. (*Erlanger, J., and H. S. Gasser, "Electrical Signs of Nervous Activity," University of Pennsylvania Press, Philadelphia, 1937.*)

After-potentials differ in the different types of fibers and they are considerably modified

by changes in the activity of the fiber or its environment:

1. *Repeated stimulation*, such as that which produces tetanus, increases the positive after-potential (Fig. 354*b* and *c*), which lasts more than 1 min. after stimulation has ceased.
2. *An increase in pH* shortens the duration of both negative and positive after-potentials. It diminishes the size of the negative after-potential and increases that of the positive one. A more marked alkalinization can suppress the negative potential, but it must be considerable to diminish the positive potential.
3. *Shifts in the electrolyte equilibrium* modify the after-potentials. Thus an increase in K^+ and a decrease in Ca^{++} depress the negative potential, and increase and shorten the positive potential. An increase in bivalent cations, or a decrease in monovalent cations, produces the opposite effects on the after-potentials.
4. *Veratrine* produces a great increase in the size and duration of the negative potential (Fig. 355), which can be larger even than the spike. Narcotics, on the contrary, depress the negative after-potential.
5. All the potentials, including the spike and resting polarization potentials, diminish in *asphyxia* and *anoxia* and recover their size when oxygen is restored. Asphyxia has a considerable effect on the after-potentials, which can disappear in anaerobiosis, but they reappear with a size two to three times the normal one when oxygen is again given.

These facts and others to be mentioned below prove that the after-potentials form part of the mechanism of the recovery of excitability. "Spike potentials can be called the messages of the nervous system; the after-potentials indicate the willingness to receive those messages for transmission."[1] All these different potentials overlap and give the composite action potential. Gasser has analyzed this potential and made a diagram of the different potentials which theoretically would give the potential registered (Fig. 356).

FIG. 356. Analysis of the combined action potential of a nerve impulse. (*Erlanger, J., and H. S. Gasser, "Electrical Signs of Nervous Activity," University of Pennsylvania Press, Philadelphia, 1937.*)

EXCITABILITY CYCLE OF THE AXON

The conduction of a nerve impulse modifies the excitability of the axon and its capacity to transmit the excitatory state. These changes can be studied by sending a second impulse at varying intervals after the first conditioning impulse and registering the action potentials, which will show the condition of the fiber.

As soon as a propagated excitatory state has been produced, *i.e.*, from the beginning of the spike, the excitability of the fiber is suppressed. The *absolute refractory period* lasts until the spike is nearly over; a second response is not obtained before the descending limb of the spike has almost reached its foot.

The recovery of excitability takes place not suddenly but gradually. Total inexcitability is followed by a period of depressed excitability.

[1] GASSER, H. L., *Harvey Lect.*, **32**, 169, 1937.

i.e., the *relative refractory period*, during which a strong stimulus produces a second spike of much smaller amplitude conducted at low speed. As the interval between the two stimuli is increased, the threshold diminishes and the voltage of the spike and the speed of conduction increase. The most excitable fibers of mammals have an absolute refractory period of 0.4 msec. and a relative refractory period of 3 msec.; but 1 msec. after the beginning of the spike a second stimulus already produces a spike of 85 to 90 per cent the voltage of the first spike (Fig. 357).

The activity of the axon is discontinuous because of its absolute refractory period. The conduction of nerve impulses should be compared to the action of a machine gun, not to that of a stream of water. The duration of the absolute refractory period puts a limit to the number of impulses that can be transmitted in a given time. Theoretically 2,500 impulses per second could be carried by the most rapid fibers, and almost 1,000 per second with a nearly normal voltage at the normal speed. The fact is that such high frequencies can be kept up only for a very short time. In physiologic conditions motor fibers transmit impulses at rates of 5 to 50 per second, exceptionally at 100 per second. Afferent fibers transmit at rates of 10 to 100 per second, and very exceptionally at higher rates up to 450 per second. There is therefore an ample margin of safety, and impulses of normal voltage are transmitted at the maximum speed at the normal frequencies of discharge of neurons and receptors.

The relative refractory period is followed by a phase of hyperexcitability during which there is a low threshold and a high speed of conduction. Later there is a period of hypoexcitability, with a high threshold and low speed of conduction. Hyperexcitability begins 3 msec. after the commencement of the spike in the most rapid fibers of mammals, reaches its peak in 3 to 10 msec., and is replaced by hypoexcitability in 12 to 18 msec. Hypoexcitability is most marked at 25 to 35 msec. and lasts altogether 60 to 80 msec. Hyperexcitability coincides with the negative after-potential and hypoexcitability with the positive after-potential. Factors that modify the after-potentials also modify the excitability of the axon. Thus veratrine increases and prolongs the negative after-potential; it also accentuates and prolongs the period of hyperexcitability.

In normal conditions, when stimuli are repeated at frequencies of 50 to 300 per second, the impulses are transmitted during the phase of hyperexcitability produced by the preceding stimulus, but repeated stimulation increases the positive after-potential and the phase of hypoexcitability. The importance of this fact is made evident in the following experiment: A nerve is submitted to moderate asphyxia so that the spike is not modified but the negative after-potential is considerably depressed. If it is then stimulated at a frequency that produced a sustained tetanus when the nerve was in normal condition, it will transmit impulses at this frequency for only a very short time.

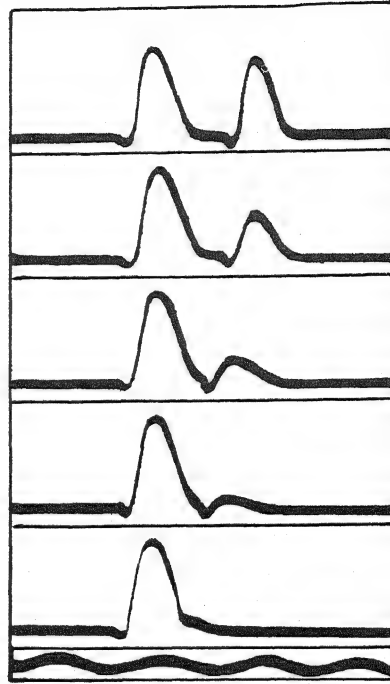



Fig. 357. Action potentials of phrenic nerve stimulated twice at varying intervals. The height of the second spike indicates the degree of recovery of the fiber, which is almost complete in the topmost record. In the lowest record the second stimulus sent during the absolute refractory period has produced a notch at the foot of the descending limb of the first spike, but no second spike. Time in msec. (*Erlanger, J., and H. S. Gasser, "Electrical Signs of Nervous Activity," University of Pennsylvania Press, Philadelphia, 1937.*)

CONDUCTION OF THE NERVE IMPULSE

The nerve takes an active part in the conduction of the nerve impulse. The stimulus does

not act like the charge of gunpowder that sends off the bullet in a cartridge, but like the spark that lights a train of gunpowder and initiates a combustion transmitted along all its length as the powder is burned up. In other words, it begins a process that is propagated along the nerve. If all the energy of the nerve impulse were originated by the stimulus and the nerve were merely a passive transmitter of this energy, as occurs with sound, the strength of the impulse would diminish as it progressed, and on arrival at its destination, it would be inversely proportional to the length of the path it has traveled. The nerve impulse does not diminish in size as it progresses along the nerve; its energy remains constant. Moreover, if it passes through a region where the properties of the axon have been depressed, its size (as given by the voltage of the spike) and its speed diminish, but on reentering a normal part of the nerve it recovers both size and speed.

Adrian[1] was the first to demonstrate this type of conduction. Later Davis and his collaborators[2] gave conclusive proof of "conduction without decrement." A length of 10 to 12 cm. of the peroneal nerve of the cat can be obtained where there is no branching off and where the nerve thus has a uniform diameter. It is placed in a box, divided into four chambers (Fig. 358). In the first chamber (*A*) the nerve is in contact with stimulating electrodes. In the second chamber (*B*) alcohol vapors are admitted to narcotize the segment of nerve it contains. Three leads (1, 2, and 3) record the action potentials at different distances from the stimulated point. The third chamber, which contains lead 4, separates the second chamber from the fourth, where air circulates and lead 5 records the action potential. Lead 6 is on a crushed portion of the nerve. The nerve can conduct impulses of normal voltage during several hours. At a given moment alcohol vapors are let into *B*, and as narcotization progresses the action potentials at 1, 2, and 3 diminish in the same degree (Fig. 359). The action potential at 5 remains normal; at 4 it may be slightly diminished, because alcohol may leak from *B* to *C*. When narcotization has reduced the action potential in *B* to about 70 per cent of its original size, the potentials at *D* also diminish a little because some of the axons are completely blocked by the alcohol in *B* and can no longer conduct impulses.

Kato and his collaborators[1] demonstrated this same fact in a single axon of a Japanese toad. A stimulus

FIG. 358. Registration of action potentials from different parts of a nerve narcotized with alcohol vapor. *A*, chamber filled with pure air where stimulating electrodes are placed; *B*, chamber filled with alcohol vapor; *C*, chamber preventing leakage of alcohol vapors from *B* to *D*; *D*, chamber filled with pure air; 1 to 6, leadoff electrodes; *G*, string galvanometer. Black segments in nerve indicate where it has been crushed. (*Davis, H., et al., Am. J. Physiol., vol. 76, p. 448, 1926.*)

[1] ADRIAN, E. D., *J. Physiol.*, **45**, 413, 1912.

[2] DAVIS, H., A. FORBES, D. BRUNSWICK, and A. M. HOPKINS, *Am. J. Physiol.*, **76**, 448, 1926.

[1] KATO, G., *Cold Spring Harbor Symposia on Quantitative Biol.*, **4**, 202, 1936.

was sent to a nerve, reduced by dissection to a single axon, and produced a maximal contraction ("all-or-nothing" law). A segment of the nerve was narcotized, and the threshold rose progressively until all conduction was suppressed. Up to that moment maximum responses were obtained. As long as the impulse could pass through the narcotized segment, it recovered its normal strength on arriving at the normal part of the axon.

These experiments can be interpreted in terms of the train of gunpowder. The narcotized segment of the nerve is like a stretch of the train that has been dampened. Combustion and transmission of the spark progress at lower speed, but once the spark reaches the dry powder it will again spread at the previous high rate. If the dampening is excessive the powder will not light and the spark will not be transmitted any further, as when in a nerve the impulses are blocked by a deeply narcotized segment.

A partial or complete block can be provoked in several ways (anesthetics, cold, etc.). A partial block can transmit impulses of low frequency but not at high frequencies because the recovery period is prolonged (Wedensky inhibition). A complete block stops the propagated impulse, but beyond the block Hodgkin has recorded a potential change (the "extrinsic potential"), which is an electrotonic potential provoked by the "spike potential" above the block. When the extrinsic potential rises to approximately 10 per cent of the spike, the impulse "jumps the block"; a spike potential is produced and is propagated along the axon beyond the block.

The nerve impulse is transmitted along the whole length of the axon including its branches at the highest speed possible, according to the "all-or-nothing" law. When an axon is stimulated simultaneously in two different points, the impulses thus started travel toward each other and when they meet cannot progress any farther because of the refractory state each one leaves in its wake.

FIG. 359. Action potentials along a nerve narcotized by alcohol vapors. (*Davis, H., et al., Am. J. Physiol., vol. 76, p. 448, 1926.*)

Saltatory conduction. In muscle fibers and unmyelinated nerve fibers local circuits spread the active state to the adjoining parts of the membrane, and conduction is apparently continuous. Lillie suggested in 1925 that in myelinated fibers excitation and the processes which maintain the propagated action potential take place only at the nodes of Ranvier; the internodes act as a unit, and conduction "jumps" from one node to the next ("saltatory" conduction).

Agents which stimulate or affect conduction, *e.g.*, electrical stimulation, blocking by electric polarization, ions, narcotics, etc., have a stronger effect at the nodes than in the internodal regions. Moreover, larger action potentials can be obtained from the nodes than from the internodal segments.[1] These facts can, how-

[1] ERLANGER, J., and E. A. BLAIR, *Am. J. Physiol.*, **110**, 289, 1934; **124**, 341, 1938; TASAKI, I., *Am. J. Physiol.*, **127**, 211, 1939; TASAKI, I., and T. TAKEUCHI, *Pflüger's Arch. f. d. ges. Physiol.*, **245**, 764, 1942; **246**, 32, 1942; TASAKI, I., and K. MIZUGUCHI, *J. Neurophysiol.*, **11**, 295, 1948; VON MÜRALT, A., "Die Signalsübermittlung in Nerven," Burkhäuser, Basel, 1946.

ever, be interpreted as due to the insulating properties of the myelin sheath which do not obtain at the nodes. Satisfactory evidence of saltatory conduction is now available. If current can enter and leave the axon only at the nodes (the myelin sheath acting as an insulator), the

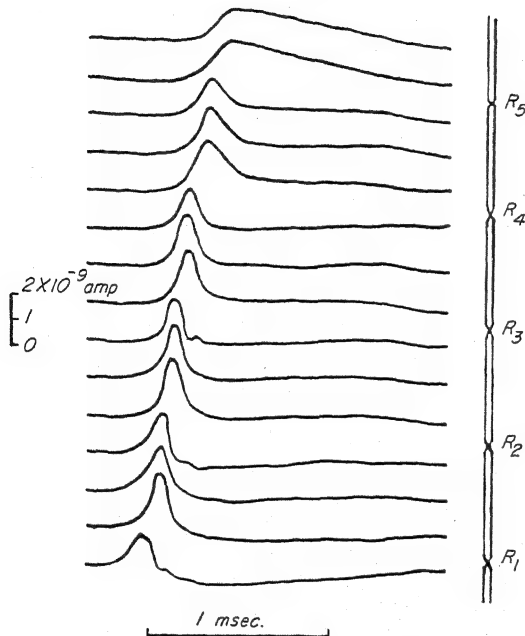



FIG. 360. Tracings of action-potential records obtained at a series of positions along one nerve fiber; diagram of nerve fiber on the right shows position where record was taken. R_1R_2, etc., nodes of Ranvier. (*Huxley, A. F., and R Stämpfli, J. Physiol., vol. 108, p. 315, 1949.*)

current in the axon at a given moment will be the same at all points of the internode. Huxley and Stämpfli[1] demonstrated this by recording in a single isolated fiber at different points of an internode the current in the external fluid, which is equal and opposite to that in the axis cylinder. Conduction time was found to be almost constant in the internode and to increase stepwise as each node was passed. The amplitude of the action potential diminished as it progressed along the internode and increased stepwise at the node (Fig. 360). The following experiment, in which conduction is blocked by increasing the external resistance, gives a very clear demonstration of the part played by the nodes: A single myelinated fiber is placed so that two adjacent nodes lie in Ringer's fluid but the internodal segment between them is in an air gap. Water evaporates on the surface of the internode, and the external conducting path becomes highly resistant. In these circumstances an impulse is not conducted through the gap, but if the nodal regions are joined by a bridge of Ringer's fluid, the impulse jumps the air gap. In myelinated fibers the local circuits go from one node to the next, the longitudinal current returning along the fluid outside the insulating myelin sheath. This mechanism increases velocity of conduction by making the local circuit act at a considerable distance ahead of the active region. Internodal length is proportional to the diameter of the fiber, and speed of conduction is proportional to fiber diameter. The fastest fibers are the thickest and have the longest internodal length.

An important objection to the theory of saltatory conduction consisted in the statement that there are no nodes in the spinal tracts made up of thick, myelinated, fast-conducting fibers. Many years ago, Cajal[1] had described "strangulations" in the fibers of tracts in the central nervous system, and lately nodes with morphologic characteristics similar to Ranvier's nodes have been described. Dyes enter the fiber at the nodes, and branching takes place only at the nodes. There is a linear relation between the diameter of the fiber and internodal spacing, thicker fibers having greater internodal distances.[2] Tasaki,[3] measuring the threshold in the dorsal tract of the spinal cord of frogs, found that at regular distances of 200 to 400 μ there were sensitive spots, while midway between them the fibers were almost insensible to stimuli four times the strength of the threshold for the sensitive places.

Fatigue. High-frequency stimulation for a sufficiently long time provokes signs of fatigue in nerves: the amplitude (voltage) of the spike potential is reduced, conduction velocity diminishes, and the refractory period is prolonged. Fatigue develops simultaneously along the whole length of the nerve, and recovery takes place at the same rate all along it. Connection with the soma is not necessary for recovery from fatigue caused by high-frequency stimulation. If the blood supply to the nerve remains intact, re-

[1] HUXLEY, A. F., and R. STÄMPFLI, *J. Physiol.*, **108**, 317, 1949.

[1] RAMON Y CAJAL, S., "Histologie du système nerveux des hommes et des vertébrés," Paris, 1911.

[2] ALLISON, A. C., and W. H. FEINDEL, *Nature, London*, **163**, 449, 1949; HESS, A., and J. Z. YOUNG, *Nature, London*, **164**, 440, 1949.

[3] TASAKI, I., *Japan. J. Physiol.*, **3**, 73, 1952.

covery is not greatly impaired following section up to the time when conduction fails owing to the process of degeneration.[1] A certain degree of recovery can take place in conditions of anoxia, but in the absence of oxygen fatigue develops more rapidly.

The mechanism of conduction of the nerve impulse. Local circuits are established by current flowing out from the region at rest into the adjacent depolarized zone. The inactive area is thus depolarized, becomes active, and exerts a similar influence on the neighboring parts of the membrane which are still at rest. The spike potential normally not only is of sufficient strength to produce this effect, but has an ample margin of safety, *i.e.*, it surpasses the threshold. For example, in the giant fiber of the squid conductance begins to increase when the spike is about halfway up the ascending limb (see Fig. 338, Chap. 66); in this case the spike has twice the threshold voltage. An even larger margin of safety (about five times) has been observed by Tasaki and Takeuchi in single-fiber preparations of the frog. If the properties of the nerve are depressed by anesthetics or in any other way, the spike is diminished and the impulse is blocked.

The role of sodium and potassium ion exchange in the genesis of the action potential and the propagated excitatory state, in this case the conduction of the nerve impulse, has been discussed in Chap. 66.

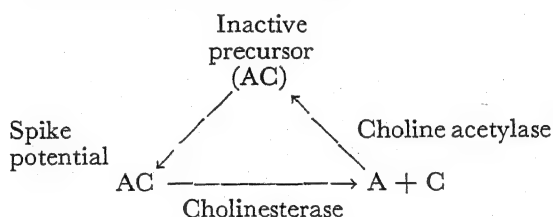
Chemical factors in conduction. Acetylcholine has been attributed an important part in the intracellular spread of excitation, similar to its role of chemical mediator at the synapse and nerve endings in myoneural junctions. This ester has been extracted from all types of nerves in many species, but the concentration differs considerably from one nerve to another. Thus the ventral spinal roots of the cat yield 12 to 22 μg per gm. of fresh weight, and sympathetic cholinergic nerves 18 to 40 μg per gm.; on the other hand, the dorsal roots and sympathetic postganglionic (adrenergic) fibers yield only very small amounts (less than 0.25 μg and 1.5 μg respectively).[2] Cholinesterase (an enzyme that accelerates the hydrolysis of acetylcholine) has also been found in nerves. In the giant axon of the squid it is concentrated near the surface, *i.e.*, in the proximity of the excitable membrane. Here again there are considerable differences in the concentration of the enzyme found in different nerves; thus the ventral roots have twice the cholinesterase activity of the dorsal roots. Another enzyme has been found which catalyzes the synthesis of acetylcholine (choline acetylase).[1] This enzyme also is found in higher concentration in the ventral roots than in the dorsal roots or the optic nerves, which have only 1 to 3 per cent the choline acetylase activity of the motor nerves.

Nachmansohn,[2] on the basis of a large number of facts collected by himself and his coworkers, supposes that acetylcholine is released from an unknown precursor by the flow of current from the inactive regions of the nerve set up by the local circuits provoked by the spike potential. The acetylcholine thus released probably acts on a "receptor protein"; it depolarizes the membrane, and a new local circuit is established. Acetylcholine is immediately destroyed by the action of cholinesterase. It is resynthesized during the recovery period by the action of choline acetylase and combined to an inactive precursor, from which it is released when the nerve is stimulated.

Inactive precursor (AC)

Spike potential: Inactive precursor (AC) → AC

Choline acetylase: A + C → Inactive precursor (AC)

Cholinesterase: AC → A + C

Liberation and inactivation of acetylcholine must take place within a very short time, because nerves can transmit more than 1,000 impulses per second. Nachmansohn and his associates have found that the concentration of cholinesterase is such that inactivation of acetylcholine could be completed in the stipulated interval.

This theory has been vigorously opposed,[3] but several of the most telling arguments against it have been answered.

[1] ERLANGER, J., and G. M. SCHOEPFLE, *Am. J. Physiol.*, **147**, 550, 1946; CAUSEY, G., and G. M. SCHOEPFLE, *J. Physiol.*, **115**, 143, 1951; CAUSEY, G., and C. J. STRATMAN, *J. Physiol.*, **120**, 373, 1953, and **123**, 234, 1954.

[2] FELDBERG, W., *Physiol. Rev.*, **25**, 596, 1945.

[1] NACHMANSOHN, D., and M. BERMAN, *J. Biol. Chem.*, **165**, 551, 1946; FELDBERG, W., and T. MANN, *J. Physiol.*, **104**, 411, 1946.

[2] NACHMANSOHN, D., *Ann. New York Acad. Sc.*, **47**, 395, 1946.

[3] GERARD, R. W., *Ann. New York Acad. Sc.*, **47**, 575, 1946.

1. The theory postulates that accumulation of ACh blocks conduction of the nerve impulse. Nerves, however, can be submerged for a long time in isotonic solutions of ACh (protected from hydrolysis by an anticholinesterase) without being blocked. This phenomenon is due to the fact that the nerve membrane is impermeable to ACh. Thus, nerve fibers were submerged in ACh labeled with N^{15}, and no labeled ACh was later found in the extruded axoplasm.[1] Acetylcholine (0.01 to 1 μg) injected into the axoplasm of giant axons of the squid prevents the development of the excitatory state under the stimulating electrode and blocks the conduction of an impulse which has arisen at some other point of the fiber. Inexcitability and block thus produced are reversible phenomena, because acetylcholine diffuses and is diluted in the axoplasm. The same effects are produced by eserine and prostigmine in 1 to 2 sec. after injecting 5 to 10 μg of these cholinesterase inhibitors. Eserine applied externally also blocks the axon, but prostigmine does not because it cannot pass through the membrane. The membrane is also impermeable to D-tubocurarine, which does not block conduction when applied externally but does so when injected into the axoplasm. This substance has only a weak anticholesterine activity, but it is very active on axonal conduction; it has been suggested that curarine probably exerts its effect by substrate competition with ACh for the "receptor protein" in the membrane.[2]
2. Cholinesterase inactivators, such as eserine, DFP, etc., according to Nachmansohn and his associates, block nerve conduction, and there is close parallelism between cholinesterase activity in the nerve and capacity to conduct impulses. This has been denied by several workers who find that cholinesterase can be inactivated in the nerve without blocking it, and that much higher concentrations of DFP and other anticholinesterases are needed to produce nerve block; in this case other enzyme systems are inhibited, *e.g.*, respiratory metabolism is depressed. It seems, however, that a considerable degree of cholinesterase activity (30 to 40 per cent of the normal value) must have remained or been recovered in experiments in which complete inactivation was supposed to have been obtained.[1]

[1] Rothenberg, M. A., D. B. Sprinson, and D. Nachmansohn, *J. Neurophysiol.*, **11**, 111, 1948.
[2] Grundfest, H., D. Nachmansohn, Y. C. Kao, and R. Chambers, *Nature, London*, **169**, 190, 1952.

TYPES OF NERVE FIBERS

The action potential of a nerve changes as it is transmitted away from the stimulated point. In the phrenic nerve the height of the spike diminishes and its duration increases, but the surface covered by the spike does not vary; therefore the total electric energy remains constant (Fig. 361). In other nerves, such as the sciatic and the saphenous, there are more complex variations. At a certain distance from the stimulated point, waves appear on the descending limb of the spike, which farther on has two more peaks (Fig. 362). This variation is due to the different speeds of conduction of the individual axons in the nerve. The impulse, therefore, does not arrive at a point some distance from the site of stimulation exactly at the same moment in all the fibers. Dispersion increases as the distance traveled is greater and the speed of conduction more varied.

Erlanger, Gasser, and Bishop,[2] classified the axons into three major groups, A, B, and C, according to the speed of conduction. Group A is made up of five subgroups: α, β, γ, δ, and ε (Fig. 363).

In a dorsal spinal root there are axons with speeds from 0.5 to 100 m. per sec.; the ventral roots have fibers with speeds of 2 to 100 m. per sec. The phrenic nerve is made up exclusively of α fibers and the sciatic of all types of fibers. The saphenous nerve has a large number of unmyelinated, slow-conducting C fibers; section of the corresponding dorsal spinal roots, peripherally to the root ganglion, causes degeneration of 80 to 90 per cent of these fibers, which are therefore afferent in nature. The remaining 10 to 20 per cent are postganglionic sympathetic axons. There are also A fibers in this nerve, of which 54 per cent belong to the α and β types, 11 per cent are γ, 25 per cent are δ, and 10 to 15 per cent are ε.

If an arbitrary value of 100 is given to the

[1] Feld, E. A., H. Grundfest, D. Nachmansohn, and M. A. Rothenberg, *J. Neurophysiol.*, **11**, 125, 1948.
[2] Erlanger, J., and H. S. Gasser, *Am. J. Physiol.*, **70**, 624, 1924; **92**, 43, 1930; Erlanger, J., H. S. Gasser, and G. H. Bishop, *Am. J. Physiol.*, **81**, 473, 1927.

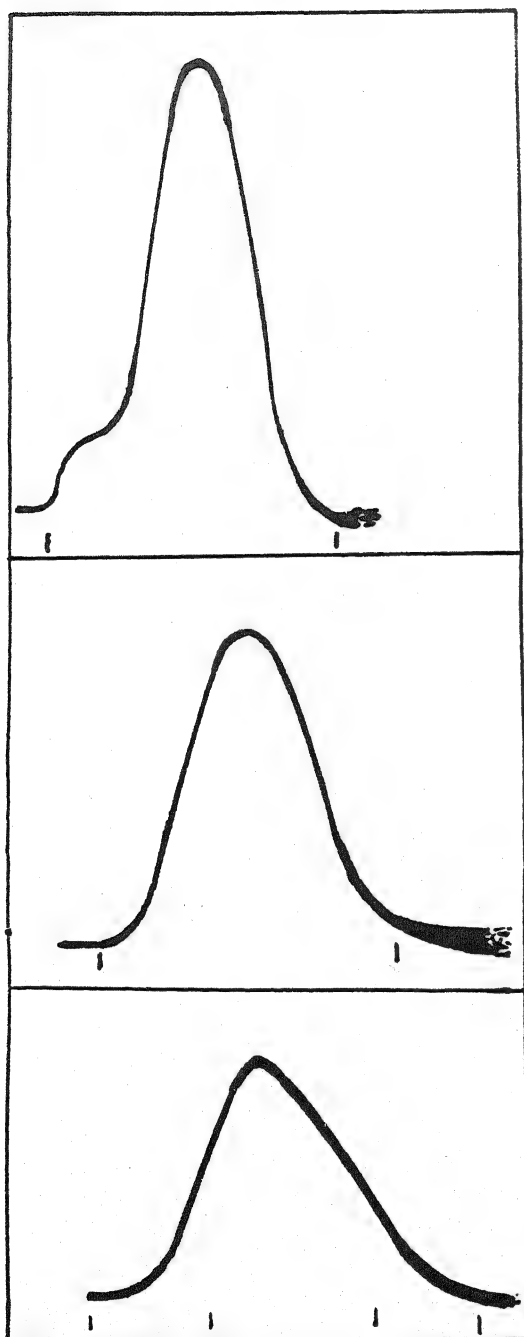



FIG. 361. Action potentials of phrenic nerve. The leads were placed at 14.5, 44, and 85 mm. from the stimulated point. (*Erlanger, J., and H. S. Gasser, "Electrical Signs of Nervous Activity," University of Pennsylvania Press, Philadelphia, 1937.*)

most rapidly conducting fibers, the following values correspond to the fastest fibers in each group: α, 100; β, 60; γ, 40; δ, 25; B, 10; C, 2. In mammals the fastest α fibers conduct at 120 m. per sec. In the bullfrog the following are typical values (in meters per second): α, 41 to

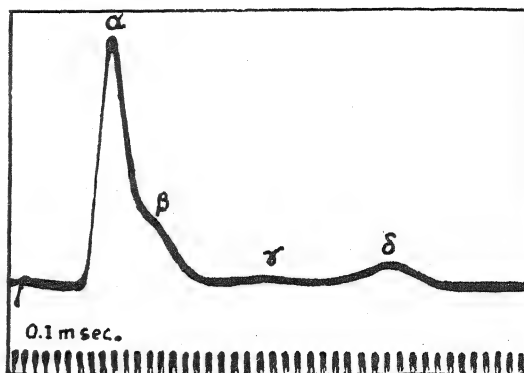



FIG. 362. Action potential of the saphenous nerve of the cat. The record was taken at 6 cm. from the stimulated point. (*Gasser, H. S., Proc. Assoc. Res. Nerv. Ment. Dis., vol. 23, p. 48, 1943.*)

47; β, 22 to 28; γ, 15 to 19; δ, 12 to 13; B, 11 to 17; C, 0.8 to 1.5. In each class there are fibers of different speeds of conduction grouped around an average, but the slowest in each class are not as rapid as the tastest in the lower class. The fibers of the various classes differ from each other in other respects besides the speed of conduction.

Internodal length, speed of conduction, spike potential, excitability, and other properties of the axons are related to their diameter. There is almost a linear relation between velocity and diameter ($V \propto D$), between spike potential and diameter ($P \propto D$), and between spike potential and velocity ($P \propto V$).[1] During fetal and postnatal growth this relation also exists. In the saphenous nerve of the cat the speed of the fastest fibers is 11 m. per sec. 4 days after birth, and 80 to 90 m. per sec. in the adult. Throughout the period of development speed of conduction increases proportionally to the growth in thickness of the fibers. Excitability is also related to fiber diameter and speed of conduction; the threshold increases as the diameter and speed of the fibers diminish. The finest fibers have long chro-

[1] LASSER, H. P., and H. GRUNDFEST, *Am. J. Physiol.*, **127**, 393, 1939; HURSCH, J. B., *Am. J. Physiol.*, **127**, 131 and 140, 1939.

naxies, but there is no simple relation between chronaxie and speed of conduction. Within a group the fastest fibers have the shortest chronaxies; there is an increase in chronaxie on passing from A to B and another from B to C. The refractory period is longer as the diameter diminishes. The period of latent addition (summation time) is 0.2 msec. in all the fibers, except in C fibers, in which it is 2.5 msec.

The action potentials differ in the different groups. All fibers have a spike and after-potentials. The A fibers in mammals have a second positive after-potential only when stimulated repeatedly; in amphibians usually there is no positive after-potential. The B fibers of mammals and C fibers of amphibians have a spike and a positive after-potential, but no negative after-potential.

B fibers are the most susceptible to asphyxia. Of the A fibers the δ and ϵ are the first to be blocked by asphyxia; the last are those of greater diameter. The C fibers are the most resistant to anoxia. Cocaine applied locally anesthetizes C fibers first, and then the myelinated fibers in the same order as asphyxia. Compression first blocks the α fibers, while considerable pressure must be applied to block the finest axons (Table 89).

Fiber type and function. The α fibers are motor fibers of striated muscles and afferent fibers of proprioceptive receptors; somatic movement is therefore served by the fastest fibers. The cutaneous receptors are innervated by fibers in the remaining A subgroups and by C fibers. No group of fibers is exclusively assigned to one type of receptor, nor is a receptor served exclusively by one type of fiber, with the exception of the proprioceptive receptors just mentioned. Tactile impulses are conducted by δ and β fibers. Pain is transmitted by C fibers, and also by myelinated fibers of several diameters. Temperature impulses are conducted by fairly thick myelinated fibers and also by very slender, possibly unmyelinated fibers (see "Conduction of sensory impulses in afferent fibers," Chap. 74). Preganglionic sympathetic fibers belong to group B. Postganglionic fibers are unmyelinated C fibers, except those of the ciliary nerve, which are myelinated and belong to group B.

FIG. 363. Fiber-size spectrum. Map of all the medullated fibers in the ninth dorsal root (above) and in the ninth ventral root (below) of the bullfrog. (*Erlanger, J., and H. S. Gasser, "Electrical Signs of Nervous Activity," University of Pennsylvania Press, Philadelphia, 1937.*)

Fibers in the spinal cord. Myelinated fibers of all sizes and unmyelinated fibers have been found in the white tracts of the spinal cord.

Propriospinal fibers, i.e., those which arise and end within the spinal cord, have been studied after cutting the cord above and below a certain

segment, so that all ascending and descending fibers arising in the spinal ganglia or higher nerve centers degenerate.[1] There are large numbers of these fibers, which are myelinated and of all sizes. In the dorsal columns small fibers, 1 μ in diameter, predominate, but in the ventrolateral columns there are many large α fibers.

Ascending fibers are also of varied sizes. The spinothalamic tracts are made up mainly of fine fibers; most of the fibers in the dorsal ascending tracts are of intermediate size and conduction velocity (70 m. per sec.). The spinocerebellar tracts are formed by thick fast-conducting (120 m. per sec.) fibers.

Descending fibers come from the frontal cortex (pyramidal tract) and from subcortical nuclei, *e.g.*, the rubrospinal, vestibulospinal, and reticulospinal tracts. The pyramidal tract is made up of fibers of many sizes; in man and the apes the largest fibers have the highest conduction velocity (α fibers), but in other animals (*e.g.* the cat) the fastest pyramidal fibers conduct at about 65 m. per sec. Rubrospinal and vestibulospinal fibers are large α fibers. The refractory period, the ratio of diameter to velocity, and other properties are similar to those of peripheral fibers.[2]

[1] TOWER, S. S., D. BODIAN, and H. HOWE, *J. Neurophysiol.*, **4**, 388, 1941.
[2] LLOYD, D. P. C., *Physiol. Rev.*, **24**, 1, 1944.

Table 89. Properties of Mammalian Nerve Fibers

Property	Group		
	A	*B*	*C*
Fiber diameter, μ	20–1	< 3	Unmyelinated
Conduction velocity, m. per sec	120–< 5	14–< 3	< 2
Spike duration, msec	0.4–0.5	1.2	2.0
Negative after-potential			
Amount, % of spike	3–5	None	3–5
Duration, msec	12–20	....	50–80
Positive after-potential			
Amount, % of spike	0.2	1.5–4.0	1.5
Duration, msec	40–60	100–300	300–< 1.000
Absolute refractory period, msec	α 0.4 δ 0.6–1.0	1.2	2.0
Period of latent addition, msec	0.2	0.2	2.5
Order of susceptibility to asphyxia	2	1	3
Order of susceptibility to cocaine	3	2	1
Order of susceptibility to pressure	1	2	3

Source: GRUNDFEST, H., *Ann. Rev. Physiol.*, **2**, 213, 1940.

THE METABOLISM OF NERVE

The structural and functional integrity of nerves is dependent on metabolic processes. A nerve separated from the organism and deprived of oxygen can, however, conduct impulses for a relatively long time. In these conditions the nerve is gradually depolarized and the membrane potential declines.[1] Activity, *i.e.*, the conduction of impulses, accelerates the process of disintegration. If the period of anoxia is not excessively prolonged, when oxygen is again available the nerve is repolarized and the membrane potential is restored. Inhibitors of respiratory metabolism, *e.g.*, cyanide, have the same effect as anoxia.

Heat production. The resting nerve of the frog at 20°C. has a heat output of 4.14×10^{-3} cal. per gm. per min., *i.e.*, 70 μcal. per gm. per sec.; an amount approximately the same as that of the resting muscle.[2] In a nitrogen atmosphere heat production diminishes gradually in the course of 3 to 4 hr. to 25 per cent of normal and remains at this level for a long time. When the nerve is again placed in oxygen, heat production increases to above the resting level; the excess heat produced is equivalent to 15 or 20 per cent of the amount that would have been put out during the period of anaerobiosis if oxygen had

[1] GERARD, R. W., *Am. J. Physiol.*, **92**, 448, 1932.
[2] BERESINA, M., *J. Physiol.*, **76**, 170, 1932.

been available. The nerve can therefore accumulate an oxygen debt, but the greater part of the energy exchanges that take place in anaerobiosis are not performed at the expense of this debt.

Oxygen consumption. The resting nerve takes up oxygen in amounts corresponding to the heat produced. Oxygen consumption varies consider .ɔly in different species and under different conditions (Table 90).

Table 90. Oxygen Consumption of Isolated Nerves

Species	*Nerve*	*Oxygen, cu. mm. per gm. per hr.*	*Season of observations*
Rana temporaria	Sciatic	23	Winter
R. temporaria	Sciatic	28	Spring
R. esculenta	Sciatic	16	Winter
R. esculenta	Sciatic	21	Spring
R. pipiens	Sciatic	37	Winter
R. pipiens	Sciatic	42	Spring
R. pipiens	Sciatic	56	Summer
Rabbit	Sciatic	280	
Dog	Sciatic	120	
Dog	Phrenic	140	
Dog	Vagus	190	
Dog	Sympathetic	210	

Source: Gerard, R. W., *Physiol. Rev.*, **12,** 469, 1932.

Amphibians have a seasonal variation, with greater oxygen consumption in times of activity (spring and summer) than in those of quiescence (hibernation). Cold diminishes heat production, and when this has fallen to 20 per cent of the normal figure, impulses are no longer conducted. Decalcifying agents (citrates) and thyroid feeding increase heat production. The Q_{10} (15 to 25°C.) is 2.2 in the frog. The RQ varies from 0.77 to 0.80.

CO_2 output. During anoxia there is a continuous output of CO_2, which is due not simply to the accumulation of acid metabolites, but to its production by oxidation of carbon. During the first 2 hr. of anoxia, CO_2 output diminishes to 50 per cent of the amount produced in aerobiosis and then drops to 10 per cent. The excess oxygen taken up during the recovery after anaerobiosis is more than can be accounted for by the CO_2 output. There seems to be an "oxygen reserve" that releases oxygen in anaerobiosis and is restored during aerobic recovery.

Metabolic changes accompanying activity. The heat production of the nerve is greater when it conducts impulses than when it is at rest.[1] At the beginning of stimulation there is a rapid increase. The methods now in use permit the detection of this increase only after many

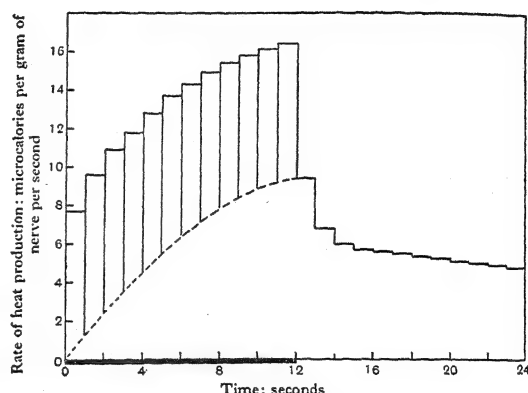



Fig. 364. Heat production of frog's nerve. The preparation was kept in oxygen during and after an almost maximal stimulus lasting 12 sec. (heavy black line). (*Hill, A. V., "Chemical Wave Transmission in Nerve," Cambridge University Press, 1932.*)

impulses have traveled down the nerve. It is not possible to say whether this excess heat is produced by the activity of the fiber or by its recovery. When stimulation ceases heat production rapidly diminishes, but during the first 20 sec. it is still considerably above the resting level; it remains above this level for 30 to 40 min. (Fig. 364).

The following phases can be distinguished in the excess heat produced by activity: (*a*) *initial heat,* produced while the nerve is conducting impulses; (*b*) *delayed heat,* produced at a high rate for a short time after stimulation has ceased, and then at a low rate for a long time.[2] The velocity of production of initial heat is 5,000 times the velocity of production of delayed heat, but as the first lasts only a short time it represents a very small part of the total excess heat. Initial heat is 3 to 4 per cent of the total heat if the stimulus lasts only a short time; *e.g.,* with 280 stimuli per second at 15°C. it is 10.5 per cent, and at 24°C. it is 8.8 per cent. When stimulation is prolonged, initial heat may be

[1] Downing, A. C., R. W. Gerard, and A. V. Hill, *Proc. Roy. Soc., London, s.B.,* **100,** 223, 1926.

[2] Hill, A. V., *Proc. Roy. Soc., London, s.B.,* **111,** 106, 1932; **113,** 345, 1933.

15 per cent of the total, because there is relatively less delayed heat.

As the frequency of stimulation increases, the heat production also increases, but only up to a certain level. The effect of an impulse is added to the effects of the preceding impulses, but the size of each impulse diminishes as the frequency increases, thus compensating the increase in number of the stimuli (Fig. 365). When the frequency is below 50 per second a steady state can be kept up indefinitely; with higher frequencies heat production gradually falls off after having reached its peak. The intervals between the impulses in this case are not long enough to allow a sufficiently complete recovery to maintain a constant heat production. The total electrical energy increases with the frequency, but the action potential of each impulse is smaller, because of incomplete recovery during the abbreviated interval. The total excess heat production due to the activity is 50 to 100 per cent that at rest; in muscle contraction the excess heat is several hundred times the resting heat production.

Oxygen consumption increases in the active nerve in relation to heat production. The excess O_2 consumption is kept up for a long time after stimulation has ceased.

In anoxia heat production by the active nerve diminishes gradually until it disappears completely. As long as heat is produced the relation between initial and delayed heat is normal, but sometimes there is relatively more delayed heat. Nerve differs from muscle in this respect because anoxia reduces considerably the delayed heat of muscle. The CO_2 output continues, as occurs in the resting nerve kept in anaerobiosis, but the oxygen debt can be greater. The asphyxiated nerve recovers its properties in oxygen and also on the addition of a H acceptor, such as *m*-dinitrobenzene.

Metabolic processes in nerve fibers. The chemical processes in nerve are similar to those of muscle. There are the same enzyme mechanisms of oxidation-reduction; several dehydrogenases and the cytochrome–cytochrome oxydase system have been found. Adenosinetriphosphatase has been found in giant axons of squids; in the sheath there is 100 times the ATP-ase activity of the axoplasm.

The isolated nerve has free glucose, which diminishes gradually and disappears in the course of several hours in the resting condition; oxygen consumption is not related to this consumption of glucose. Glycogen, hexosephosphate, and lactic acid do not show considerable changes during rest. The stimulated nerve consumes glucose at the same rate as the nerve at rest. In anoxia lactic acid accumulates in amounts equivalent to the glucose that has disappeared; on adding glucose, more lactic acid is formed. In aerobic recovery glycogen is not resynthesized, nor is lactic acid oxidized; therefore, although glycolysis occurs in the nerve, the Pasteur-Meyerhof cycle does not take place.

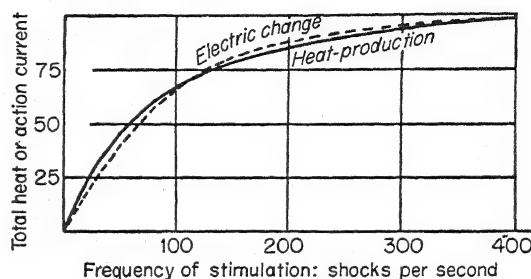



Fig. 365. Relation between total heat per second (solid line) or total action current (broken line) and frequency of stimulation. (*Gerard, Hill, and Zotterman.*)

Phosphocreatine is disintegrated during rest in anaerobiosis, with a consequent increase in phosphate; at first this takes place at a fast rate, and later more slowly. Phosphocreatine is resynthesized during aerobic recovery. Inhibition of glycolysis by monoiodacetate accelerates phosphocreatine disintegration; but an increase in glycolysis does not spare phosphocreatine. In a stimulated nerve these same processes take place at a faster rate.

The resting nerve sets free ammonia both in aerobiosis and in anaerobiosis, and stimulation doubles its production. Deamination of oxidized proteins might be the source of NH_3, but this is doubtful, because it is still produced in anaerobiosis when oxidation of protein by the oxygen reserve has ceased. Another probable source is the disintegration of adenylpyrophosphate, which occurs in nerve.

The significance of the acetylcholine cycle has already been considered.

Relatively large amounts of thiamine have been found in myelinated nerves. There is a progressive decrease of this vitamin as degeneration progresses after a nerve has been cut. Deficiency of thiamine causes degenerative

processes in peripheral nerves and in the white tracts of the spinal cord. Vitamin C (ascorbic acid) has also been found in nerves, but its significance is as yet unknown.

The trophic influence of the perikaryon is probably a metabolic one, *i.e.*, it is implemented by chemical reactions. An axon separated from its cell body does not suffer from dystrophia due to inactivity, because stimulation does not delay degeneration but accelerates it, probably by speeding up the consumption of substances normally supplied by the perikaryon. A nerve can be compressed during several weeks without provoking degeneration, although all impulses are blocked. Normal conduction is reestablished in 1 to 2 hr. after compression has ceased.

The facts known about the metabolism of nerve at rest and during activity cannot yet be coordinated into a consistent theory explaining the metabolic and energy processes responsible for the maintenance of the specific properties and function of the nerve fiber, *i.e.*, excitability and conduction of impulses.

BIBLIOGRAPHY

ADRIAN, E. D., "The Mechanism of Nervous Action," Oxford, New York, 1932.

BISHOP, G. H., The Relation of Bioelectric Potentials to Cell Functioning, *Ann. Rev. Physiol.*, **3**, 1, 1941.

BISHOP, G. H., Nerve and Synaptic Conduction, *Ann. Rev. Physiol.*, **8**, 355, 1946.

BREMER, F., Nerve and Synaptic Conduction *Ann Rev. Physiol.*, **9**, 457, 1947.

BRONK, D. W., and F. BRINK, Bioelectric Study of Excitation and Response of Nerve, *Ann. Rev. Physiol.*, **1**, 385, 1939.

DAVIS, H., The Conduction of the Nerve Impulse, *Physiol. Rev.*, **6**, 547, 1926.

ECCLES, J. C., Conduction and Synaptic Transmission in the Nervous System, *Ann. Rev. Physiol.*, **10**, 93, 1948.

ERLANGER, J., and H. S. GASSER, "Electrical Signs of Nervous Activity," University of Pennsylvania Press, Philadelphia, 1937.

GERARD, R. W., Nerve Metabolism, *Physiol. Rev.*, **12**, 469, 1932.

GERARD, R. W., The Metabolism of Brain and Nerve, *Ann. Rev. Biochem.*, **6**, 419, 1937.

GERARD, R. W., Electrophysiology, *Ann. Rev. Physiol.*, **4**, 329, 1942.

GRUNDFEST, H., Bioelectric Potentials, *Ann. Rev. Physiol.*, **2**, 213, 1940.

GRUNDFEST, H., Bioelectric Potentials in the Nervous System and in Muscle, *Ann. Rev. Physiol.*, **9**, 477, 1947.

LLOYD, D. P. C., Conduction and Synaptic Transmission in the Nervous System, *Ann. Rev. Physiol.*, **6**, 349, 1944.

LORENTE DE NÓ, R., "A Study in Nerve Physiology," Stud. Rockefeller Inst. Med. Res., vols. 131 and 132, New York, 1947.

MCCOUCH, G. P., Conduction and Synaptic Transmission in the Nervous System, *Ann. Rev. Physiol.*, **7**, 455, 1945.

NACHMANSOHN, D., in G. Pincus and K. V. Thimann, "The Hormones," Academic Press, New York, 1950.

RENSHAW, B., Nerve and Synaptic Transmission, *Ann. Rev. Physiol.*, **5**, 253, 1943.

WHITTAKER, V. P., Specificity, Mode of Action, and Distribution of Cholinesterase, *Physiol. Rev.*, **31**, 12, 1951.

CHAPTER 69

Conduction in the Nerve Center

THE ELEMENTARY REFLEX

Transmission of excitation from one neuron to another is conditioned by two processes, the *central excitatory state* and the *central inhibitory state*,[1] which interact to coordinate the response of effectors to stimuli applied on the receptors. Conduction of an impulse along a neuron is to a certain extent an automatic process, because once the threshold for the propagated excitatory state has been attained, the whole neuron is excited. New possibilities arise in the synapse: excitation can be transmitted to a large number of neurons or to only a few; resting neurons can be excited, or the activity of those in a state of excitation can be suppressed. Synaptic transmission is, therefore, of fundamental importance in nervous activity, and the reflex is the primary act of integration, the functional unit of the nervous system, because in every reflex at least one synapse is involved.

The reflex arc. The structural base of the reflex is the reflex arc, made up of (*a*) the *receptor*, an organ that receives, and is excited by, the stimulus; (*b*) the *afferent neuron*, which transmits the excitatory state to the center; (*c*) the *synapse;* (*d*) the *efferent* or *motor neuron*, which transmits the excitation to the periphery, or responds by ceasing to discharge impulses, if the afferent volley is of an inhibitory, instead of an excitatory, nature; (*e*) the *effector*, in which the nerve impulse evokes a response adapted to the stimulus (Fig. 318).

The *two-neuron reflex arc*, with only one synapse, was first described by Cajal, who showed that some of the dorsal-root fibers end directly on the motor cells of the ventral horn; many of these fibers are collaterals of fibers in the dorsal column. Dorsal-root section and subsequent observation of the degenerating synaptic terminals (knobs, *boutons*, or end-feet) have given further proof of the existence of these two-neuron arcs. Lloyd's[1] work on conduction and synaptic transmission of stretch reflexes (monosynaptic reflexes) has confirmed the conclusions of the histologists. There are relatively few two-neuron arcs; in the lumbar segments of the cat, less than 10 per cent of the afferent fibers end directly on a motor neuron. These are α fibers of low threshold and high speed of conduction. They innervate receptors situated in the muscle and end on spinal motor neurons, which discharge impulses provoking a myotatic or stretch reflex in the muscle in which the receptor is located.

The great majority of reflex arcs, including those of all the cutaneous reflexes, are *multineuron reflex arcs*, having one or more *connector* or *internuncial neurons* between the afferent and efferent neurons.

The reflex arc described as a chain of neurons is a useful abstraction, but it does not really exist. An afferent neuron innervates more than one receptor and makes connections with more than one motor or internuncial neuron, and the latter receive endings from more than one centripetal or internuncial neuron. Thus when a single receptor is stimulated many neurons may enter into activity and provoke the response of several effectors. This is known as the *principle of divergence*. On the other hand, stimulation of different receptors may provoke a response through the same motor neuron which acts as a final common path. This is known as the *principle of convergence*.

Stimulation of a receptor or of an afferent

[1] Sherrington introduced the term *central excitatory state* (c.e.s.) to describe the enduring excitatory condition set up by a centripetal volley, without reference to its precise nature. The *central inhibitory state* (c.i.s.) is an enduring effect produced by a centripetal volley antagonistic to the c.e.s.

[1] LLOYD, D. P. C., *J. Neurophysiol.*, **6**, 316, 1943.

nerve does not produce contraction or relaxation in a single muscle, but in a group of muscles, which act synergically to produce a movement. The nervous system of higher animals is not organized in anatomical segments but in movement patterns and in visceral reactions. "The simplest spinal reflex 'thinks,' so to say, in movements, not in muscles" (Jackson).[1]

Fig. 366. Receptive field of the scratch reflex in the spinal dog. The figures give the relative strength of the threshold stimulus in different parts of the field. *lr*, last rib. (*Sherrington, C. S.*, "*The Integrative Action of the Nervous System,*" *Yale University Press, New Haven, 1926, p. 221.*)

Elementary reflexes. All vertebrates show certain typical reflexes, which consist in stereotyped responses to specific stimuli applied to definite receptive fields. One group of reflexes produces movement of the whole or part of the body; they are called *phasic reflexes*. An important group of these are defensive reactions, the object of which is to remove the nocuous stimulus or withdraw the body from its influence. The simplest is the *flexion reflex*, *i.e.*, flexion of the limb on which the painful stimulus acts. Another group of reflexes serves to maintain the equilibrium of the body and its posture; these are the *postural reflexes*. The simplest of these is Sherrington's *stretch*, or *myotatic*, *reflex*, which consists in the localized contraction of a muscle in response to stretching. If there is an excessive pull, the shortening ceases and is replaced by a lengthening of the muscle; this *lengthening reaction* is also of reflex origin but is produced by central inhibition, not by excitation. The *extensor thrust* is also a postural reflex; it consists in a sharp and brief extension of the limb on the application of pressure to the plantar surface, or on stretching the foot muscles by separating the toes, as occurs when the weight of the body is placed on a foot. This reflex is not restricted to one muscle or to part of a muscle, as in the case of the stretch reflex, but involves all the extensors; it forms part of Magnus's positive supporting reaction (see "Postural reflexes," Chap. 81). If the stimulus is such that it causes pain, the response is a flexion reflex instead of an extensor thrust.

These reflexes have the afferent and efferent limbs of the arc in the same spinal segment; they are *segmental*, or *short*, *reflexes*. In others, the neurons of the arc occupy several segments; they are *intersegmental*, or *long*, *reflexes*. The scratch reflex observed in the dog is an example of this last type of reflex. It consists in rhythmic alternate flexion and extension of the hind limb when a stimulus is applied to a saddle-shaped area of the skin extending from between the shoulders to the sacral region (Fig. 366).

Recent work tends to show a fundamental similarity between the processes of intraneuronal and synaptic transmission, but conduction in the reflex arc differs in many respects from conduction in the nerve. The main characteristics of conduction in the reflex arc will now be discussed.

THE LAW OF FORWARD DIRECTION OF CONDUCTION

When a nerve fiber is stimulated, at whatever point the stimulus acts, the propagated excita-

[1] Quoted by Sherrington, C. S., *Brain*, **54**, 1, 1931.

tory state spreads in both directions and the whole neuron is excited. On the other hand, transmission of excitation through a synapse takes place in only one direction and always in the same sense, *i.e.*, from the axon endings of one neuron to the dendrites or perikaryon (the soma) of the neuron with which these endings connect. This fundamental difference between intraneuronal and interneuronal conduction has been called by Sherrington the "law of forward direction of conduction." Irreciprocal conduction at the synapse is one of the most important physiologic foundations of the neuron doctrine.[1]

Bell-Magendie law. Bell (1811) observed that mechanical stimulation of the ventral spinal roots provoked muscular contraction, an effect that did not follow stimulation of the dorsal roots. A short time afterward (1822), Magendie demonstrated that section of the dorsal roots suppressed sensibility but not motility in the corresponding segments, and he confirmed Bell's finding that stimulation of the ventral root has motor effects. This fact is now known as the Bell-Magendie law. The ventral roots do not carry centripetal impulses; an experimental antidromic[2] stimulus is not transmitted to the dorsal roots or to the spinal tracts. On the other hand, antidromic conduction, *i.e.*, transmission of centrifugal impulses, has been demonstrated in the dorsal roots.

Antidromic conduction. The following are some of the most important facts demonstrating antidromic conduction in the dorsal roots:

1. Stimulation of the peripheral end of a cut dorsal root or of a cutaneous nerve provokes vasodilatation, called by Bayliss "antidromic vasodilatation." The phenomenon has been observed in the cat's paw, rabbit's ear, etc., and in man (Foerster).
2. Stimulation of the central end of the cut vagus nerve produces reflex hypotension, which persists after removal of the sympathetic, *i.e.*, of vasomotor nerves (Bayliss).
3. Mustard oil, or any other irritating substance, when applied on the skin, provokes vasodilatation and other symptoms of inflammation. This effect is suppressed by anesthetizing the skin with cocaine, or by cutting the dorsal roots peripherally to the ganglion, or by denervation of the skin, after the nerves have degenerated, but not before (Bruce).
4. Toennies[1] has registered (by means of a cathode-ray oscillograph) impulses traveling in the dorsal roots after stimulation of a cutaneous nerve (the internal saphenous) in the cat. A flexion reflex is provoked and at the same time there is a centrifugal discharge of impulses in the ipsilateral and contralateral dorsal roots. About one-third of the α fibers and some δ fibers are involved in this activity, which has all the characteristics of a reflex discharge.[2] The central reflex time is 3.5 to 4.5 msec. in the ipsilateral and 4.2 to 6.5 msec. in the contralateral fibers. Summation, facilitation, afterdischarge, and inhibition are observed. Asphyxia quickly suppresses the reflex and antidromic discharges.

The neurons that conduct these antidromic impulses have not yet been identified. There are two possibilities: (*a*) impulses are conducted antidromically by the sensory ganglionic neurons; (*b*) impulses are conducted by spinal neurons, which send out their axons in the dorsal roots to the periphery, or to the spinal ganglion, where they connect synaptically with the sensory neurons. Section of the dorsal root between the ganglion and the spinal cord does not produce degeneration in all the fibers of the proximal end (the end near the cord); 12 to 15 days after the operation intact fibers are found, even when the ganglion has been removed. Some doubt has been cast on the significance of this observation by the facts that the dorsal roots send out collaterals to the neighboring roots and that there are recurrent fibers of the ventral roots that join the dorsal roots; the undegenerated fibers found after section of a dorsal root could belong to either of these groups. No synapses have been found in the spinal ganglia, and there is a 1 : 1 proportion between the cells in a ganglion and the fibers it sends to the spinal cord. It is therefore difficult to admit that there are any fibers in the root other than those coming from the sensory ganglion cells.

Bruce's experiments, just referred to, and those of Lewis on the "triple response" of the blood vessels of the skin following trauma have suggested the "axonic reflex" hypothesis. According to this hypothesis, an

[1] Reverse or antidromic conduction will be discussed later.

[2] Stimulation of a ventral root causes an excitatory state propagated to the periphery, where it produces motor effects, and toward the spinal motor centers, *i.e.*, in a direction opposite (antidromic) to that of physiologic impulses.

[1] Toennies, J. F., *J. Neurophysiol.*, **1**, 378, 1938; **2**, 515, 1938.

[2] The following discussion will be understood better after reading the rest of the chapter, in which the characteristics of a reflex discharge are considered.

axon branches out at the periphery and innervates the skin and the blood vessels. A stimulus applied to the cutaneous branch or to the nerve trunk spreads throughout the neuron and therefore to the vascular branch, thus producing vasodilatation. This is not a true reflex—there is no *interneuronal* conduction—but a pseudoreflex produced by *intraneuronal* spread of excitation. The long latency (several seconds) and duration (several minutes) of antidromic vasodilatation suggested that a chemical mediator released at the nerve endings was responsible for this phenomenon.[1] Acetylcholine (Wybauw), histamine (Kwiatkowski), and an unknown substance, perhaps identical with the central synaptic transmitter,[2] have been considered as the chemical mediator. Experiments made on antidromic vasodilatation in the denervated rabbit's ear have shown that vasodilatation provoked by acetylcholine is suppressed by atropine and potentiated by eserine, but that atropine has no influence on antidromic vasodilatation and eserine diminishes it. Small doses of histamine produce vasodilatation, which is suppressed together with other effects of histamine by antihistamine drugs, such as mepyramine, which have no effect on antidromic vasodilatation. Acetylcholine or histamine cannot therefore act as mediators of antidromic vasodilatation in this case.[3] In others, however, histamine is released by antidromic impulses. Thus, antidromic stimulation of the dorsal roots corresponding to the hind limb in dogs provokes vasodilatation in the skin which is suppressed by antihistaminics (mepyramine, antistine, pyribenzamine).[4]

Fibers conducting Toennies' antidromic impulses are not recurrent collaterals of neighboring roots, because no peripheral degeneration was seen after cutting the rootlets in which they had been registered. A large number of neurons are involved in this phenomenon. Also Toennies has shown that afferent impulses, such as are produced by lightly tapping the skin, diminish considerably when they are preceded by an antidromic discharge. These facts suggest that the same neurons take part in afferent and antidromic conduction. Antidromic discharges may be due to interaction between neighboring fibers in the dorsal spinal tracts[5] facilitated by the abnormal condition of the cord produced by experimental conditions. Interaction between peripheral myelinic fibers made hyperexcitable has been observed,[1] also in the recently cut and cooled spinal cord.[2]

There is as yet no completely satisfactory explanation of "antidromic" conduction in the dorsal roots.

[1] LEWIS, T., and H. M. MARVIN, *J. Physiol.*, **62**, 19P, 1926.

[2] HELLAUER, H. F., and K. UMRATH, *Pflüger's Arch. f. d. ges. Physiol.*, **249**, 619, 1948.

[3] HOLTON, P., and W. L. M. PERRY, *J. Physiol.*, **114**, 240, 1951.

[4] IBRAHIM, F. A. D., *Quart. J. Exper. Physiol.*, **36**, 189, 1951.

[5] BARRON, D. H., *J. Neurophysiol.*, **3**, 403, 1940.

THE LATENT PERIOD

There is a certain delay between stimulation of a receptor or an afferent nerve and the reflex motor response. This delay is called the "latent period" of the reflex. It is due to the time taken for the impulse to travel from the receptor, or the point on the afferent fiber which has been stimulated, to the nerve center, for activation of the center, transmission along the motor fiber, and finally activation of the effector. The time left after deducting peripheral conduction times of the afferent and efferent paths is known as the *central reflex time.*

Eccles and Sherrington[3] have measured the central reflex time of the flexion reflex (Table 91).

Table 91. Latent Period of the Flexion Reflex

	Msec.
Afferent nerve, conduction time, 13.8 cm. at 31.6 m. per sec	4.4
Efferent nerve, conduction time, 19.5 cm. at 93 m. per sec	2.1
Total peripheral conduction time	6.5
Total latent period	10.4
Central reflex time	3.9

In the spinal cat it was usually 3 to 5.5 msec., for a reflex provoked by a single stimulus. The latent period was usually shorter after a strong stimulus than after a weak one. This difference was due exclusively to reduction of the central reflex time, because the conduction of the impulses along the nerve fibers is an "all-or-nothing" phenomenon. When two successive stimuli were sent at an optimum interval, the central reflex time was reduced to 0.5 msec. The central reflex time is the sum of the *central conduction time, i.e.*, the time taken by the impulse to travel along the fibers in the nerve center, and the *synaptic delay*, the time spent in transmitting the excitatory state from one neuron to another.

[1] ROSENBLUETH, A., *Am. J. Physiol.*, **132**, 119, 1941.

[2] RENSHAW, B., and P. S. THERMAN, *Am. J. Physiol.*, **133**, 96, 1941.

[3] ECCLES, J. C., and C. S. SHERRINGTON, *Proc. Royal Soc., London, s.B.*, **107**, 511, 1931.

Synaptic delay. Lorente de Nó[1] has measured the synaptic delay in the motor nucleus of the third cranial nerve. He stimulated the motor neurons directly by means of a fine electrode introduced into the gray substance of the nucleus and the presynaptic fibers in the longitudinal bundle (Fig. 367). The maximum delay was found to be 0.8 to 0.9 msec., but under certain conditions it could be reduced to 0.5 to 0.6 msec. Renshaw[2] has measured this delay in the motor nuclei of the spinal cord, using a fine electrode placed in the ventral horns and recording the action potentials in the ventral root. A single stimulus of sufficient strength produces two spikes in the action potential. The first spike is due to the discharge of the motor neurons stimulated directly; the second is less synchronized and is due to stimulation of the motor neurons through the internuncial neurons. The interval between the spikes varies from 0.6 to 1 msec., but can be reduced to 0.1 to 0.3 msec. by increasing the strength of the stimulus or sending two stimuli separated by an adequate interval. The synaptic delay in other neurons has also been found to be 0.5 to 1 msec.

The synaptic delay in sympathetic ganglia is much greater than in the central nervous system. Stimulation of preganglionic fibers to the superior cervical ganglion provokes discharges from the ganglion cells after a synaptic delay of about 2 msec. when the fibers stimulated belong to the Aδ group, which end on large neurons, and of about 5 msec. when they are finer fibers of the B group, which end on medium-sized neurons.

[1] LORENTE DE NÓ, R., *Am. J. Physiol.*, **111**, 272, 1935.
[2] RENSHAW, B., *J. Neurophysiol.*, **3**, 373, 1940.

The so-called "synaptic delay" includes the conduction time along the fine terminal branches of the axons. This conduction time must amount to something, because the branches do not follow a straight line in their course, and it is well known that conduction velocity is low in slender fibers. The synaptic delay, in a strict sense, must therefore be even shorter than the times given.

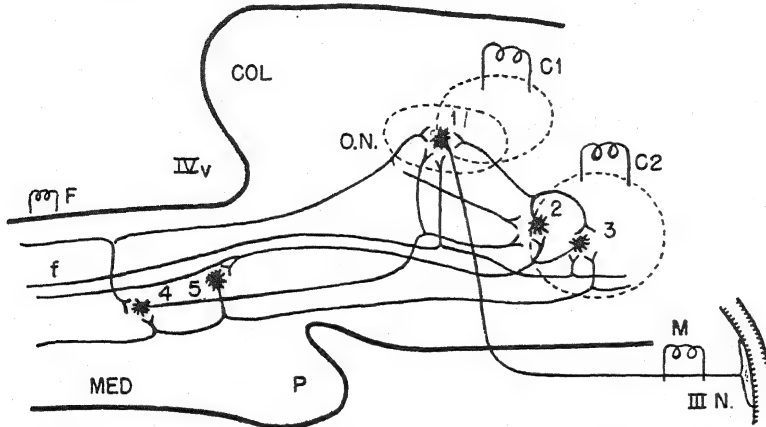



FIG. 367. Diagram of Lorente de Nó's experiment for determination of synaptic delay. *O.N.*, motor nucleus of the third nerve; *COL*, colliculi; *MED*, medulla oblongata; *P*, pons; III *N.*, oculomotor nerve; IV*v*, fourth ventricle; *C1*, *C2*, *M*, and *F*, stimulating electrodes; *f*, fibers of longitudinal bundle; 1, motor neuron of the oculomotor nerve; 2, 3, 4, and 5, internuncial neurons. (*Lorente de Nó, R., Am. J. Physiol., vol. 111, p. 272, 1935.*)

Lloyd[1] distinguishes the *synaptic delay* from the *nuclear delay*. The latter consists in the interval between the arrival at the synapse of the *first* presynaptic impulse and the beginning of postsynaptic activity. Synaptic delay is the interval between the arrival at the synapse of the *last* presynaptic impulse necessary to provoke transmission and the beginning of postsynaptic activity. When all the impulses arrive almost simultaneously, as occurs in the nucleus of the third cranial nerve in the conditions of Lorente de Nó's experiment, nuclear and synaptic delays have approximately the same duration, and the synaptic delay is fairly constant. On the other hand, when the impulses arrive asynchronously, dispersed in time owing to the different speeds of conduction and different lengths of the presynaptic paths, as occurs in spinal reflexes, the nuclear delay is much greater than the synaptic

[1] LLOYD, D. P. C., *Physiol. Rev.*, **24**, 1, 1944.

delay. The distinction between nuclear and synaptic delays is important, because in most cases neurons are excited by a barrage of asynchronous impulses from internuncial neurons. In the interval between the first and last impulses that arrive at the surface of the neuron, the central excitatory state is built up.

Fig. 368. Diagram of motor neuron pool. Only two afferent fibers, *A* and *B*, have been drawn. In *a*, impulses traveling along *A* stimulate motor neurons α and β; those along *B* stimulate motor neurons β and γ. On simultaneous excitation of *A* and *B*, the β neurons are occluded. The subliminal fringe has been eliminated so as not to complicate the figure. In *b*, impulses from *A* stimulate motor neurons α above the threshold, so that they discharge impulses, and motor neurons β subliminally; impulses from *B* stimulate motor neurous γ which discharge, and motor neurons β subliminally. On simultaneous stimulation of *A* and *B* the subliminal c.e.s. provoked in the β neurons summates and these also discharge. (*After Sherrington, C. S., Proc. Roy. Soc., London, s.B., vol. 105, p. 332, 1929.*)

THE PRINCIPLE OF CONVERGENCE. THE FINAL COMMON PATH

Afferent fibers are more numerous than efferent fibers. Thus in the cat the average number of all the fibers in the dorsal roots of one side is 535,000, and that of the ventral fibers is 115,000; the proportion of ventral to dorsal fibers is 1:4.65.[1] Similar figures have been observed in other species. The predominance of sensory fibers is even more marked in the cranial nerves, owing to the great cephalic receptors of vision, hearing, taste, and smell. All sensory fibers can provoke reflexes; therefore several of them must *converge* to the same motor center. The anatomic structure of the neuron, with many synapses on the soma and only one axon to discharge the impulses, makes for convergence.

Many physiologic facts concur with these structural features to prove that impulses conducted by different sensory neurons converge to a *final common path*. The reflex contraction of a muscle, *e.g.*, the tibialis anticus, can be obtained as part of the flexion reflex, by stimulation of several afferent nerves (Table 92). The maximum reflex tension obtained in each case is less than the maximum tension obtained by tetanic stimulation of the motor nerves. No afferent path stimulates maximally all the neurons innervating the muscle; the center is distributed, or fractionated, among several sensory paths. On the other hand, the sum of all the reflex tensions gives a much higher figure than the maximal tension obtained by a motor tetanus. Therefore each motor unit must respond to more than one afferent neuron.

Occlusion. The principle of convergence is also demonstrated by the phenomenon known as "occlusion." Simultaneous stimulation of two afferent nerves, concurring in a reflex, produces a smaller reflex tension than the sum of the tensions obtained by the separate stimulation

[1] Duncan, D., and L. L. Keper, *J. Comp. Neurol.*, **64**, 303, 1936; **68**, 479, 1938; Holmes, F. W., and H. A. Dewenport, *J. Comp. Neurol.*, **73**, 1, 1940.

of each nerve. For example, simultaneous stimulation of both plantar nerves provokes a reflex tension of 1.81 kg. in the tibialis anticus; if each is stimulated separately, reflex tensions developed are 1.57 and 1.58 kg. respectively; the difference between the sum of these tensions

Table 92. Fractionation of the Motor Center; Occlusion (Tibialis anticus. Maximum motor tension 2,160 gm.)

Afferent nerve stimulated	*Tension of maximal reflex tetanus, gm.*	*Reflex tension expressed as percentage of maximal motor tetanus*
Internal saphenous	800	32
Superficial obturator	165	6.7
Deep obturator	400	16
Nerve to quadriceps and sartorius	1,190	44
Musculocutaneous branch of peroneus	1,700	69
External plantar	1,240	50
Internal plantar	1,330	54
Small sciatic	680	28
Hamstring	565	23
Nerve of sural triceps	300	12
Sum	8,370	

Source: Creed *et al.*

(1.58 + 1.57 = 3.15) and that obtained by simultaneous stimulation (3.15 − 1.81 = 1.34) gives the amount of occlusion. The "occluded" neurons respond maximally to both afferent nerves; occlusion is due to overlap of the areas in the motor center stimulated by the afferent neurons (Fig. 368*a*). Occlusion is conditioned by (*a*) the structural organization of the nerve centers, *i.e.*, the anatomic proximity of the afferent path endings; (*b*) physiologic factors, such as frequency of stimulation (occlusion increases with frequency of stimulation), fatigue, spinal shock, and anesthetics, which reduce the spread of excitation and therefore diminish occlusion.

SUMMATION AND FACILITATION

The central excitatory state produced by a centripetal volley resembles the local excitatory state produced by a stimulus applied to a nerve or muscle fiber: (*a*) it takes some time to develop and decays following an exponential curve; (*b*) it spreads decrementally from the locus of excitation; (*c*) if it covers a sufficient area and attains a critical level (the threshold), a propagated impulse is fired off, which is transmitted over the whole postsynaptic neuron and is followed by a refractory state. For this threshold to be attained it is necessary that several synaptic endings, placed near each other on the cell surface, should discharge within a very short time (see "Synaptic transmission," below). The latent period, extension, and intensity of the c.e.s. are dependent on the strength of the afferent volley. If a weak volley, which produces only subliminal excitation, is repeated, a reflex response may be evoked; the effects of the second volley are added to those of the first, and the passage of the impulse provoked by the second stimulus is facilitated. This is known as *temporal summation.* Lorente de Nó's work on the oculomotor nucleus has deprived this type of summation of all importance in the building up of c.e.s. at an individual synapse (see further on). If two weak volleys converge to the same neuron pool along different paths, there may be summation of the subliminal c.e.s. elicited by each one of them, and a reflex response will be observed. This is known as *spatial summation.*

Subliminal fringes. A strong afferent volley impinging on a motor neuron pool stimulates a few of the motor neurons maximally and others submaximally. The active neurons are surrounded by others excited subliminally, which form a "subliminal fringe" around the "discharging zone" (Fig. 368*b*). Neurons situated at a greater distance are not affected by the afferent volley and remain quiescent (Fig. 369). If the strength of the stimulus is increased, a larger number of fibers in the afferent nerve are excited, the area of c.e.s. is enlarged, quiescent neurons are "recruited" into the subliminal fringe, and neurons from the subliminal fringe enter into the discharging zone. Lloyd[1] has shown that the afferent volley must attain considerable size before the motor neurons pass from the subliminal fringe to the discharging zone. Further progressive increase in the strength of the afferent volley causes a rapid increase in the discharging zone ands ubliminal fringe, and a linear relation is established between the two.

When an afferent nerve is stimulated, both types of summation (temporal and spatial) may occur, because the nerve is made up of many fibers with different conduction velocities, which

[1] Lloyd, D. P. C., *J. Neurophysiol.*, **6**, 111, 1943.

converge to a neuron pool by paths of different length.

Spatial summation. (Fig. 368*b*.) For spatial summation to occur, the convergent volleys must impinge on the neuron pool simultaneously or with only a short interval between them. If

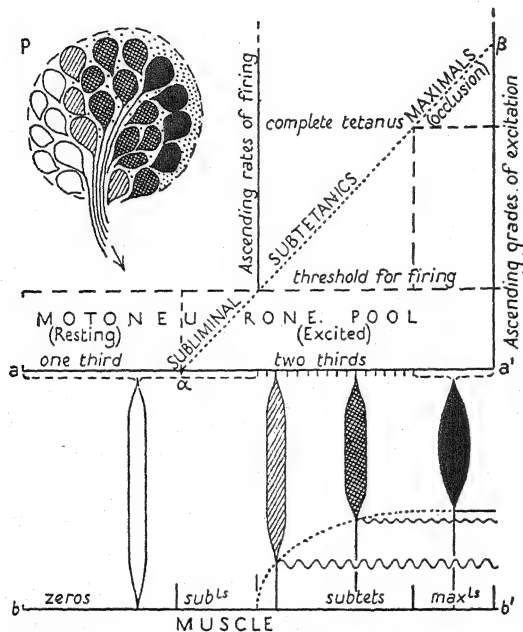



FIG. 369. Diagram of activation of a motor neuron pool. (*Creed et al., "Reflex Activity of the Spinal Cord," Oxford University Press, New ork, 1942, p. 122.*)

the interval is progressively increased, the reflex response diminishes and will eventually disappear if it is the result of summation in a subliminal fringe (Fig. 370). The nucleus of the third cranial nerve is particularly suitable for demonstration of the phenomena of summation.[1] The motor neurons can be stimulated directly or through the longitudinal bundle (Fig. 367), and synchronized volleys can be sent with little or no temporal dispersion. Two volleys sent by different paths produce a maximum effect when they are simultaneous; the effect diminishes as the interval between the volleys increases, and there is no summation when the interval is 0.5 msec. or more. In sympathetic ganglia spatial summation has been demonstrated by stimulation of convergent preganglionic fibers. In this case also, the convergent volleys must arrive at the ganglion simultaneously or with a minimum interval for summation to occur.

Temporal summation. This type of summation consists in the accumulation of the effects of successive impulses which arrive at a neuron along the same afferent fiber. Lorente de Nó has examined the possibilities of temporal summation in the nucleus of the third cranial nerve. The summation interval for successive stimuli applied directly to the neurons was from 0.1 to 0.2 msec., equivalent to the absolute refractory period of the afferent fiber. Therefore temporal summation at an individual synapse does not play a part in building up c.e.s.

In sympathetic ganglia, conditions differ. A period of hyperexcitability follows the refractory state of the ganglionic neuron; therefore a second subliminal volley that arrives at the synapse during this period of supernormality may provoke a response, *i.e.*, the first volley has facilitated the effect of the second.

The latent period of the response to the second stimulus can be reduced to a minimum equivalent to the synaptic delay (0.3 to 0.4 msec.), by adjusting the strength of the first stimulus so that it is just below the threshold, and applying the second stimulus after

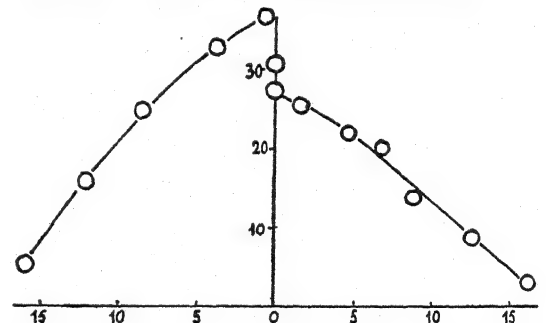



FIG. 370. Reflex tension in relation to interval between stimuli. Response of tibialis anticus measured in grams of tension (ordinates) when two nerves are stimulated at varying intervals. To the right of zero, the lateral gastrocnemius nerve is stimulated first; to the left of zero, the median gastrocnemius nerve is stimulated first. Abscissas, time in milliseconds. (*Creed et al., "Reflex Activity of the Spinal Cord," Oxford University Press, New York, 1942, p. 32.*)

an interval that is neither too short nor too long (6 to 8 msec.). The second volley thus arrives at the motor neuron pool when the c.e.s. produced by the first is at its peak (Fig. 371). The central reflex time taken from the application of the first stimulus to the reflex response corresponds to the nuclear delay.

[1] LORENTE DE NÓ, R., *Am. J. Physiol.*, **111**, 272 and 283, 1935; **112**, 595, 1935; **113**, 505 and 524, 1935.

Internuncial "barrage." Spatial summation of subliminal c.e.s. plays an indispensable part in intercellular transmission. In Lorente de Nó's experiments on the nucleus of the third cranial nerve, it was possible to excite the motor cells by a single stimulus only when other paths converging to the nucleus were intact and transmitted impulses that maintained a certain level of subliminal c.e.s. Internuncial neurons have thus two functions: (*a*) they form links in the reflex arc and serve as channels of distribution of afferent impulses; (*b*) they maintain a subliminal c.e.s. by constant bombardment of the cells on which their axons end. Two fundamental types of circuits assure this continuous action of the neurons.[1] In one type (Fig. 372*M*), the impulses arrive at the center temporally and spatially dispersed, because they travel along paths in which different numbers of synapses are interposed. In the second type (Fig. 372*C*), the neuron chains form a closed circuit. In this case the recurrent impulse is retarded by the synaptic delay sufficiently to assure its arrival after the end of the refractory period left by the previous impulse. The circuit of excitation is interrupted by the summation of hypoexcitability left by successive impulses, and the neurons no longer respond to the afferent impulses.

FIG. 371. Diagrammatic representation of c.e.s. in motor neuron pool (ordinates) plotted against time (abscissas). Above: a single volley sets up a c.e.s. which attains the threshold T_1 after a latent period of 1. Below: the second stimulus is applied at varying intervals after the first. The latent period is shortest (3) when the interval is such that the second volley reaches the center at the peak of the c.e.s. elicited by the first. (*After Eccles, J. C., and C. S. Sherrington, Proc. Roy. Soc., London, s.B., vol. 107, p. 511, 1931.*)

AFTER-DISCHARGE

The motor response of a reflex persists for a much longer time after the afferent volleys have ceased than the time taken by the last stimulus to travel along the reflex arc. In some cases the maximum effect is observed after stimulation of the receptor has ceased. After-discharge may be due, in part, to stimulation of receptors by the effects of the reflex, but this is not the only and necessary mechanism of this phenomenon, as it is observed in the flexion reflex despite deafferentation of the limb by section of the corresponding dorsal roots. After-discharge is produced by repetitive firing of the motor center.[2]

Repetitive stimulation is the rule in a normal nervous system. The receptors send, along the afferent neurons, repetitive impulses that vary in frequency and duration according to the strength of the stimulus. Internuncial neurons keep up a continuous "barrage" of impulses on the neurons placed "downstream" in the chain, and the motor neurons discharge repetitively on the effectors. Prolonged activity of the internuncial neurons (interneuron reverberation) is, therefore, one of the mechanisms that cause after-discharge.

FIG. 372. Types of circuits: *M*, multiple chain; *C*, closed circuit of neurons. (*Lorente de Nó, R., J. Neurophysiol., vol. 1, p. 207, 1938.*)

[1] LORENTE DE NÓ, R., *J. Neurophysiol.*, **1**, 207, 1938.
[2] *Ibid.*, **2**, 402, 1939.

The following observations demonstrate repetitive firing of motor neurons due to activity of internuncial neurons:

1. Repeated stimulation of a motor nerve at a lower frequency than 50 per second produces an incomplete tetanus, the oscillations of which have the same rhythm as the stimulus. Repeated stimulation of an afferent nerve at the same frequency also produces incomplete tetanus, but the oscillations of the electromyogram have a greater frequency than the rhythm of stimulation. The motor neurons discharge several times for every afferent impulse. Thus contraction is smoother, the oscillations are less marked, and a complete tetanus is obtained with lower frequencies of stimulation than those necessary to produce it by excitation of the motor nerve.
2. A single stimulus produces action potentials in the afferent roots which last 1 to 2 msec. The electrical activity of the ventral root when thus reflexly aroused can last 25 msec. An electrical record of the spinal cord taken at the same time shows that the activity of the motor neurons lasts as long as the internuncial neurons are active.

After-discharge can, however, be observed where there is no activity of internuncial neurons. Thus, in sympathetic ganglia (which have no internuncial neurons) strong stimulation at relatively high frequency (60 per second) of the preganglionic fibers is followed by postganglionic discharges that persist at a falling rate for some time (27 msec.) after the presynaptic volleys have ceased.[1] Bremer and his associates[2] have examined, in the spinal frog and toad, the intense prolonged reflex after-discharge of "tonic" muscles. They found that it was more marked after deafferentation and transection of the cord, *i.e.*, when only the lower lumbosacral segments remained active. In these conditions reflex reinforcement was not possible, and interneuron reverberation was reduced to a minimum. This work coincides with that of Rosenblueth and his associates[3] in concluding that self-sustained activity of neuron aggregates is an important mechanism of after-discharge.

[1] Larrabee, M. G., and D. W. Bronk, *Federation Proc.*, **5**, 60, 1946.

[2] Bremer, F., V. Bonnet, and J. Moldaver, *Arch. internat. de physiol.*, **52**, 215, 1942.

[3] Rosenblueth, A., J. G. Ramos, and W. B. Cannon, *Arch. Inst. Cardiol. México*, **15**, 401, 1945.

INHIBITION

Suppression or prevention of activity in an effector by the action of the nervous system is known as inhibition. More than a hundred years ago (1845), the Webers demonstrated that stimulation of the vagus nerve stopped the heartbeat. For the first time suppression of activity, instead of an increase in it, was seen to follow stimulation of a nerve. Later the inhibitory effects of the splanchnics and other visceral nerves were observed, and the inhibitory nerves of invertebrates were discovered. In all these instances inhibition takes place in the innervated structure; there is *peripheral inhibition.*[1]

In somatic movements, suppression of activity (*i.e.*, inhibition) in the antagonistic muscles is as important as stimulation of the activity of the protagonists for accurate and smooth performance of movements. There are, however, no inhibitory nerves to somatic muscles in vertebrates. Centrifugal stimulation of a nerve passing into a muscle provokes contraction only; there is no peripheral inhibition. Inhibition of somatic activity is exclusively a central process, the counterpart of the central excitatory state, which Sherrington has called the "central inhibitory state" (c.i.s.). This central inhibitory state is made evident only by suppression of the central excitatory state.

Characteristics of inhibition. Inhibitory effects can take place with great rapidity, and the inhibited muscle relaxes at the same speed at which it contracts. The central latency of inhibition, however, is as variable as that of the central excitatory state.

Inhibition is graded in intensity. The degree of inhibition depends on the relative proportions of c.e.s. and c.i.s. This can be demonstrated in the crossed-extension reflex in the following way (Fig. 373): A cutaneous nerve of a limb is stimulated centripetally. Ipsilateral flexion and contralateral extension are thus provoked. Contraction of an extensor (*e.g.*, the quadriceps) in the contralateral limb is registered. When the reflex is fully developed, a cutaneous nerve of the extended limb is stimulated, and the quadriceps will be seen to relax immediately. The degree of relaxation is much greater than that observed by merely suppressing stimulation of the cutaneous nerve on the flexed limb for the same

[1] The mechanism of peripheral inhibition will be discussed in Chap. 84, because in the vertebrates all inhibitory nerves are visceral nerves.

length of time as the inhibitory stimulus has been applied. The amount of inhibition, *i.e.*, the number of motor units inactivated, will depend on the strength of the inhibitory stimulus and the amount of potential c.e.s. present. If the inhibitory stimulus is weak, no apparent effect may be registered; if, on the other hand, it is strong and the c.e.s. present is weak, complete inhibition may occur. After-discharges are particularly sensitive to inhibition, and a weak inhibitory stimulus, which has no effect while the excitatory stimulus is active, may quickly suppress contraction due to after-discharge.

Convergence and summation of inhibitory impulses are observed in conditions similar to those which produce convergence and summation of excitatory impulses. A weak stimulus, which does not provoke manifest inhibition, may nevertheless produce latent or subliminal inhibition which can be made evident by repetitive stimulation, or by stimulation of afferent paths that converge to the same center. A motor center is fractionated among the afferent paths with respect to inhibition in the same way as with respect to excitation. Moreover, a subliminal fringe of inhibition surrounds the pool of completely inhibited neurons. Occlusion is observed in the inhibited neurons just as in active neurons.

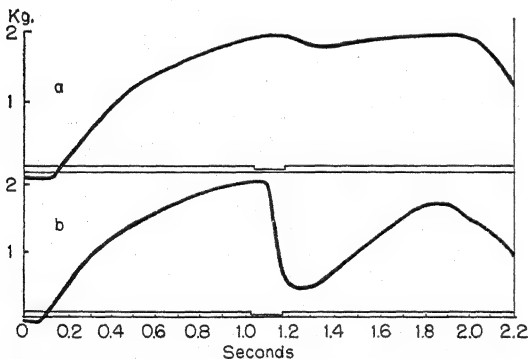



Fig. 373. Crossed-extension reflex in the decerebrate cat. Reflex contraction of the quadriceps in response to stimulation of the peroneopopliteal nerve of the opposite side at the rate of 48 stimuli per second. In *a*, three stimuli are dropped. In *b*, three stimuli are applied to an ipsilateral cutaneous nerve, and inhibition is produced. (*Liddell, E. G. T., and C. S. Sherrington, Proc. Roy. Soc., London, s.B., vol. 95, p. 90, 1923.*)

The time course of inhibition is as variable as that of the c.e.s.. but inhibition can last for a considerable time. Thus inhibition of the knee jerk (contraction of the quadriceps) may last for 2 sec. after a single shock to an ipsilateral cutaneous nerve. In the experiment recorded in Fig. 374, reflex contraction of the tibialis anticus was inhibited by a conditioning stimulus. The inhibitory effect increased to a maximum, which was reached between 30 and 80 msec. Then it diminished, but even after 200 msec. there was a considerable degree of inhibition. In this experiment several afferent fibers of a multineuron reflex were stimulated; therefore the inhibitory impulses arrived at the motor center by several paths dispersed in time and space, and a c.i.s. was kept up for a long time.

Fig. 374. Inhibition of the reflex contraction of the tibialis anticus by a conditioning stimulus. Ordinate, tension in grams developed by the muscle. Abscissa, interval in msec. between the stimuli. The arrow marks the tension developed when no conditioning stimulus preceded the second stimulus. (*Eccles, J. C., and C. S. Sherrington, Proc. Roy. Soc., London, s.B., vol. 107, p. 548, 1931.*)

The exact nature of the process of inhibition is not known, but several observations have shown there are two mechanisms of inhibition, which have been called *indirect inhibition* and *direct inhibition.*

Indirect inhibition. Conduction of an impulse (the propagated excitatory state) is followed by a prolonged subnormal phase during which the threshold is raised; this phase coincides with the positive after-potential. Its long duration facilitates summation of successive subnormal phases produced by repetitive stimulation, and the neuron becomes progressively less excitable. If subnormality is sufficiently pronounced, the neuron will not respond and im-

pulses that impinge on it will appear to be inhibited. This is known as indirect inhibition.

The simplest case in which this occurs is the so-called "silent period" of the knee jerk. This is a two-neuron reflex; the arc is formed by large α afferent fibers from the dorsal roots which end directly on motor neurons. If the electrical activity of the rectus femoris is registered, small asynchronous oscillations due to "tonic" contraction will be recorded. A tap on the knee tendon provokes extension of the leg accompanied by a short series of ample oscillations in the electrical record. Electrical activity is then suspended for a few milliseconds (the "silent period") before the asynchronous waves reappear (Fig. 375). The "silent period" can be registered with greater accuracy in a single motor unit activated by rhythmic tonic impulses. A single stimulus applied to an afferent nerve provokes a reflex discharge interpolated in the "tonic" rhythm, followed by a period of complete inactivity. The silent period is longer when the "reflex" impulse comes close to the preceding "tonic" impulse. An antidromic volley has the same effect as the conditioning afferent stimulus; therefore the silent period is due to a change in the excitability of the motor neuron. Moreover, summation of subnormal phases and positive potentials corresponding to successive antidromic volleys prolongs the silent period.

Fig. 375. Electromyogram of the rectus femoris muscle of a human subject. At the signal the tendon is tapped, and a knee jerk follows. There is a silent period, during which the electrical activity of the muscle is suspended; when it is renewed the oscillations are more ample. Time in milliseconds. (*Lindsley, Am. J. Physiol., vol. 109, p. 181, 1934.*)

This type of inhibition, due to the previous activity of the neuron, has also been demonstrated in the internuncial neurons of the spinal cord by means of the flexion reflex. Two afferent volleys are sent along the same cutaneous nerve. Response to the second volley will be facilitated or inhibited according to the relative strength of the volleys and the interval between them. Inhibition of the second stimulus is most marked when a weak stimulus is followed by a strong one after an interval of at least 16 msec. in order to avoid facilitation. Hughes and Gasser[1] have recorded the electrical activity of the spinal neurons following stimulation of afferent fibers, comparing the results with the motor response. The nature of the response could be predicted by the record of the spinal-cord potential. If the positive after-potential was small, facilitation occurred; if the positive after-potential had a high voltage, inhibition in proportion to the size of the positive after-potential was observed.

Direct inhibition. Indirect inhibition is the consequence of previous activity in the neuron, but in certain cases inhibition occurs in neurons that can be considered as having been quiescent. There must, therefore, be another mechanism of inhibition. Lloyd[2] has demonstrated what is known as direct inhibition, in the two-neuron reflex arc. Internuncial neurons played no part in this case; inhibition took place at the synapse of the afferent dorsal root and the spinal motor neuron. A segmental reflex was evoked by stimulating the first sacral dorsal root, and the motor discharge in the corresponding ventral root was recorded. A conditioning volley was sent along the sixth lumbar dorsal root of the same side. When the volleys were synchronized, or the conditioning volley preceded the test volley by a short interval, inhibition was observed. Maximum inhibition was obtained with an interval of 0.7 msec. between the volleys; afterward inhibition decayed exponentially, but with a threshold volley the inhibitory effect lasted up to 7 or 8 msec. If the conditioning volley arrived at the synapse after the test volley, no inhibition was observed; therefore, in this case, after the c.e.s. was established it could not be suppressed by a subsequent inhibitory impulse.

Afferent inhibitory fibers are large, low-

[1] Hughes, J., and H. S. Gasser, *Am. J. Physiol.*, **108**, 307, 1934.

[2] Lloyd, D. P. C., *J. Neurophysiol.*, **4**, 184, 1941.

threshold, high-speed α fibers which come from proprioceptive receptors in a muscle and end on the motor neurons of its antagonists.[1]

FATIGUE

Prolonged repetitive stimulation is followed by progressive decrease in the response. This is not inhibition but fatigue. Inhibition is the rapid suppression of activity by another stimulus, while in fatigue the afferent path loses control of the motor neurons without the concurrence of another stimulus.

Phasic reflexes are easily fatigued; in the spinal cat the flexion reflex begins to diminish in less than 1 min. Postural reflexes are less susceptible to fatigue.

The motor unit is not the site of fatigue; it can give a full response to direct stimulation, or to a reflex impulse that arrives through another afferent path, after it has ceased to respond to a given impulse. Fatigue does not reside in the receptor or the afferent nerve but in the nerve centers. The nature of fatigue is still unknown, but it seems to be related to the intense metabolic activity of the centers.

METABOLIC ACTIVITY OF THE NERVE CENTERS

Nerve centers are particularly susceptible to lack of oxygen. A reflex center ceases to function after a few minutes of anoxia, while nerve fibers conduct impulses during a considerable time in an oxygen-free atmosphere. Heymans and Bouckaerts[2] have examined the relative sensitiveness to anoxia of the different nerve centers. The higher centers, particularly the cortex, were found to be very sensitive to lack of oxygen; ischemia of 5 min. duration produced irreversible damage, and in some cases death. Recovery was possible if the blood supply was interrupted for a shorter time, but even then blindness often occurred, owing to the great susceptibility of the striate cortex (occipital lobe). Recovery of subcortical centers governing postural reflexes was observed if ischemia did not last more than 10 to 15 min.; decerebrate rigidity (see Chap. 81) was caused by more prolonged ischemia. Respiratory, vasomotor, and pupillary reflexes were definitely abolished by cutting off the blood supply for 30 to 60 min., and all spinal reflexes were permanently suppressed if ischemia lasted over 1 hr.

Oxygen consumption of slices of cerebral cortex, determined *in vitro*, is in the order of 5,000 cu. mm. per gm. per min. during the first few minutes, and later falls to 2,000 cu. mm. These figures have no great value, because nervous tissue is a delicate structure that suffers considerable damage by this experimental procedure, but they show that the neuron soma has much higher oxygen consumption than the axon, which in similar conditions takes up only 40 cu. mm. per gm.

Susceptibility to drugs. The high metabolic rate of the nerve centers is associated with their susceptibility to certain drugs such as strychnine, anesthetics, and narcotics (barbiturates). Low concentrations of these drugs, which have little or no effect on the axons, depress or totally suppress reflex activity.

ELECTRICAL PHENOMENA OF SYNAPTIC TRANSMISSION

Transmission of an impulse from one neuron to another (synaptic transmission) is accompanied by electrical phenomena. These have been recorded in autonomic ganglia (stellate, superior cervical, ciliate, lumbar, inferior mesenteric), in the spinal cord following stimulation of receptors or afferent fibers, in motor nuclei such

i
0.5 mv.
ii
a
i
0.5 mv.
ii
0.2 mv.
iii
b

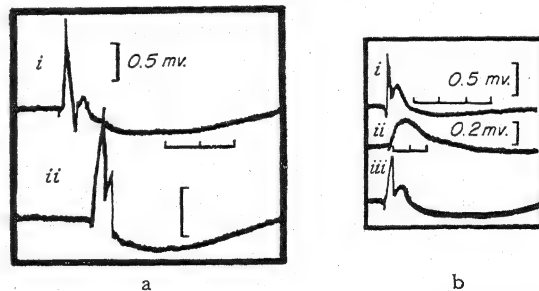



Fig. 376. (*a*) Action potentials set up in ganglionic neurons by single presynaptic volleys. (*b*) Action potentials set up by single presynaptic volleys in (*i*) normal and (*ii*) curarized ganglion, and (*iii*) after soaking out tubocurarine. Slower time course is due to lower temperature of preparation (isolated superior cervical ganglion of cat). (*Eccles, R. M., J. Physiol., vol. 117, p. 181, 1952.*)

as the motor nucleus of the third nerve or those of the anterior horn of the spinal cord, and by recording potentials of a single spinal motor neuron with an intracellular electrode.

Electrical phenomena in autonomic ganglia. (Fig. 376.) Stimulation of the pregan-

[1] *Ibid.*, **9**, 421 and 439, 1946.
[2] Heymans, C., and J. J. Bouckaerts, *Compt. rend. Soc. de biol.*, **119**, 424, 1935.

glionic fibers in a curarized ganglion evokes a potential—*synaptic potential*[1]—similar to the end-plate potential (see Chap. 67). It reaches its summit in 10 to 15 msec. in the cat's stellate ganglion *in situ*, and decays exponentially with a time to half decay of 60 to 90 msec. This potential spreads with decrement along the postsynaptic neuron, *i.e.*, it has the characteristics of electrotonic potentials. The size of the synaptic potential is related to the concentration of curare; as the drug diffuses out and its effect diminishes, the synaptic potential increases and may reach the threshold of the propagated excitatory state, in which case a postsynaptic impulse with spike potential is fired off. This occurs when the amplitude of the synaptic potential rises to a critical level, which is 5 to 12 per cent of the spike amplitude in the cat stellate ganglion *in situ*, but may be more in other cases, *e.g.*, 22 to 29 per cent in the isolated superior cervical ganglion.[2] Synaptic potentials evoked by repetitive stimulation of adequate frequency summate, and a potential may be built up to the threshold level. The synaptic potential is not modified significantly by eserine when it is evoked by a single presynaptic volley, but this drug prolongs it considerably when it is produced by repetitive stimulation; release of acetylcholine, therefore, plays a part in its production (see pages 851–853).

In a noncurarized ganglion the synaptic potential is covered by the spike potential, but in certain conditions it can be seen as an inflection in the rising limb of the spike. Usually there are several postsynaptic spikes because of differences in the speed of conduction of preganglionic fibers[3] and asynchronic activation of the ganglionic neurons. The number of spikes increases with the strength of the stimulus; a sufficiently strong stimulus evokes not only one or two spikes with a relatively short latency, but also one or even two others with long latency due to stimulation of nonmyelinated preganglionic fibers.[4]

Ganglionic spikes are followed by a *negative afterpotential*, due in part to the synaptic potential which outlasts the spike, and in part to an electrical phenomenon similar to the negative afterpotential of the nerve. There is also a *positive afterpotential*, analogous to the corresponding positive waves in nerves. Repetitive stimulation is followed by larger and prolonged negative and positive afterpotentials.

[1] Eccles, J. C., *J. Physiol.*, **101**, 465, 1943.

[2] Eccles, R. M., *J. Physiol.*, **117**, 181, 1952.

[3] Bishop, G. H., and P. Heinbecker, *Am. J. Physiol.*, **94**, 170, 1930; **100**, 519, 1932.

[4] Rosenblueth, A., and F. A. Simeone, *Am. J. Physiol.*, **122**, 688, 1938.

Posttetanic potentiation is observed after repetitive presynaptic stimulation (1 to 30 sec.); single volleys evoke larger spikes, and the positive wave is at first depressed, then enhanced. Repetitive stimulation may be followed by afterdischarge, *i.e.*, persistent activity of the ganglionic neurons after stimulation has ceased.[1] Afterdischarge is easily observed in eserinized ganglia.

Motorneuron potentials. (Fig. 377.) Electric phenomena of motor neurons have been recorded by means of fine insulated-wire electrodes inserted into a motor nucleus, but a more precise analysis has been made by means of glass electrodes introduced into a single motor-neuron of the anterior horn of the spinal cord.[2] The average resting potential was found to be 70 mv., but probably the maximum figure (80 mv.) is nearer the true value because experimental conditions tend to diminish it. An antidromic impulse (stimulation of the ventral root) evokes a spike potential of 95 mv. (average) which reverses the resting potential by 25 mv.; the true value is probably nearer 110 to 120 mv., the maximum figures recorded. The spike lasts 1 msec., twice the duration of the spike of large motor fibers; it is followed by a negative afterpotential, which after several milliseconds is reversed to a prolonged positive afterpotential reaching a maximum of 3 to 6 per cent of the spike in 10 to 15 msec., and lasting for about 100 msec.

An orthodromic impulse evoked by stimulating a muscle nerve, in order to provoke a monosynaptic reflex, is followed by a synaptic potential after a latency of 0.6 msec. This potential reaches its summit in 0.6 to 1 msec. after the onset, then decays exponentially, falling to half its value in 2.5 to 3.5 msec., and is later reversed to a small positive afterpotential which is increased by repetitive stimulation. The antidromic spike rises at a rate of 300 to 500 volts/sec. (half the rate at which the motor fiber spike rises), but the synaptic potential rises slowly. If

[1] Larrabee, M. G., and D. M. Bronk, *Proc. Soc. Exper. Biol. & Med.*, **38**, 921, 1938; *J. Neurophysiol.*, **10**, 139, 1947.

[2] Brock, L. G., J. S. Coombs, and J. C. Eccles, *J. Physiol.*, **117**, 431, 1952.

the rate of rise is 50 volts/sec., it reaches the threshold and the spike potential of the propagated impulse appears, in which the synaptic potential is marked by an inflection in the ascending limb. The critical voltage for this is 6 to 14 mv., 8 to 15 per cent of the spike.

two or more stimuli are applied to the afferent fibers at adequate intervals, the synaptic potentials evoked by each impulse summate, and a potential can be built up which reaches the threshold and fires off a postsynaptic impulse with a spike potential (temporal summation).

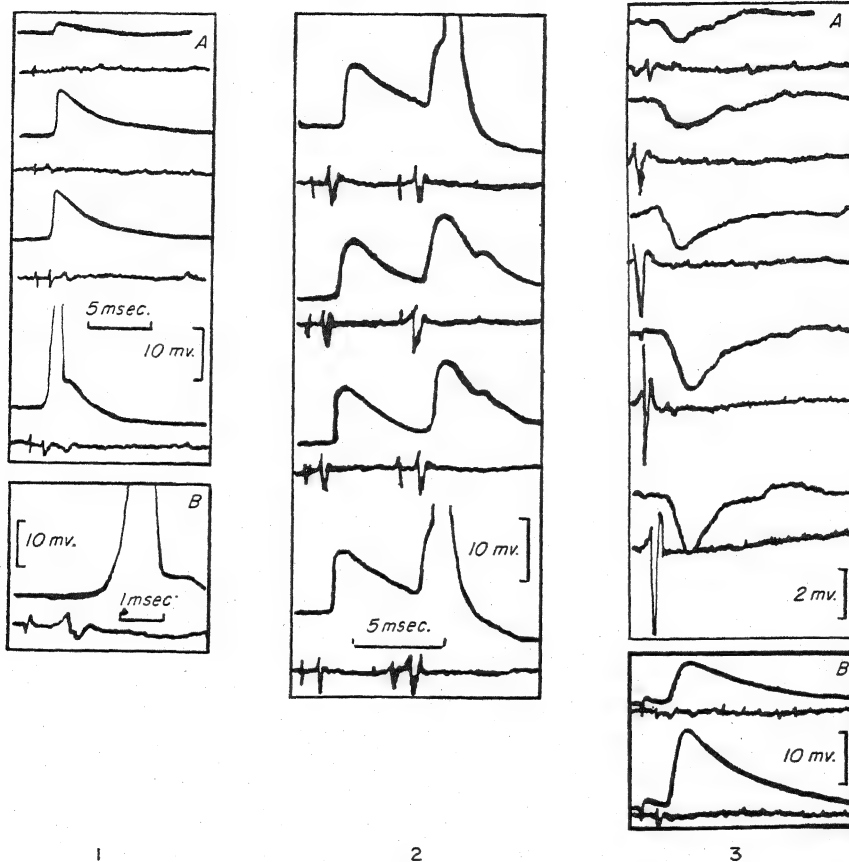



Fig. 377. 1.*A*, intracellular potentials set up in spinal motor neuron by afferent volleys; synaptic potentials increase with size of volley; the largest sets up a spike. *B*, faster record shows spike arising from initial synaptic potential. Lower tracings in all the records correspond to dorsal root potential, negativity downward.

2. Intracellular potentials set up in spinal motor neuron by two afferent volleys from the same fibers. Summation of synaptic potentials that set up a spike occurred at the shortest interval (upper record). The intervals in the second and fourth record, which are equal, are just critical.

3. Intracellular potentials of spinal motor neuron responding with synaptic potentials to afferent volleys from semitendinosus biceps nerve (*B*). Volleys of increasing size (from above downward) from quadriceps (antagonist of semitendinosus and biceps) nerve evoke positive potentials in the same motor neuron as in (*B*). (*Brock, L. G., J. S. Coombs, and J. C. Eccles, J. Physiol., vol. 117, p. 431, 1952.*)

The synaptic potential increases with the strength of the stimulus until it reaches the threshold. This increase is due to activation of a larger number of end knobs so that depolarization spreads over a wider area of the motor-neuron membrane (spatial summation). When There is evidence suggesting that under each end knob maximum depolarization occurs but that a sufficient area of the motor-neuron membrane must be activated before a self-propagated process is initiated, *i.e.*, before an impulse with a spike potential will spread throughout the neuron.

If the electrode is moved out of the motor neuron so that its tip is placed in the intercellular space, the postsynaptic potentials are modified, but there is no change in the presynaptic potentials. Therefore the latter when recorded by an intracellular electrode are extracellular in respect of the presynaptic neuron. This is definite proof of discontinuity at the synapse and crucial demonstration of Cajal's neuron doctrine (Fig. 378).

A

1mv.

B

10mv.

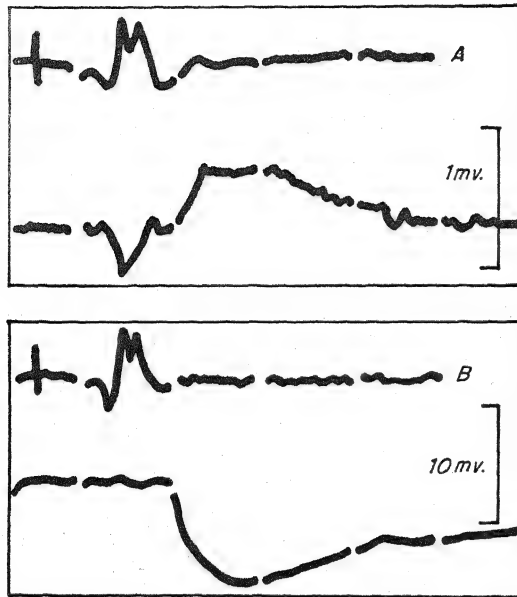



FIG. 378. Potentials recorded in response to afferent volley; upper record, dorsal-root potentials (negativity upward); lower record, motor-neuron potential (negativity downward). *A*, extracellular record; *B*, intracellular record. Root potentials do not vary. Motor-neuron potentials recorded intracellularly are reversed and increased with respect to the extracellularly recorded potentials. (*Brock, L. G., J. S. Coombs, and J. C. Eccles, J. Physiol., vol. 117, p. 431, 1952.*)

When a motor neuron which responds with a synaptic potential to stimulation of a given muscle nerve, *e.g.*, the biceps-semitendinosus nerve, receives an impulse set up by stimulation of the nerve of a muscle which is an antagonist of the former, *e.g.*, the quadriceps, it responds not by a synaptic potential but by increased negativity with respect to earth, *i.e.*, by an increase in polarization. This potential reaches a maximum of 1 to 2 mv. in 1 msec., then decays to half its value in 2 msec. A stimulus which provokes direct inhibition, therefore, is followed by a state of anelectrotonus in the motor neuron. This state is not produced by potentials outside the cell, but is an active response of the motor neuron to the inhibitory volley. The significance of this fact will be discussed later.

Spinal-cord potentials. Spinal-cord potentials can be recorded by placing one electrode on the dorsal surface near the dorsal root stimulated and the other on the ventral surface of the same segment or of an upper segment. Stimulation of a dorsal root evokes a spike followed by a second negative potential of lower voltage and greater duration; finally there is a prolonged positive potential of low voltage (Fig. 379). The spike is due to intraspinal fibers of dorsal ganglion neurons, and when two roots are stimulated simultaneously there is summation of their respective spikes. The second negative potential has been considered as the sum of the spikes of internuncial neurons,[1] the lower voltage and greater duration being due to temporal dispersion in the excitation of these neurons. When two dorsal roots are stimulated, the second negative potential is not the sum of the respective individual root potentials, as occurs with the spikes; it is only slightly larger than that of a single root because afferent fibers of both roots converge to the same internuncial neurons, some of which are occluded. The positive potentials of the neurons first excited may also contribute to lower the voltage of the second negative potential.

When a volley of impulses, or even a single impulse, enters the spinal cord by a dorsal root, slow, prolonged negative potentials can be recorded from the same or neighboring dorsal roots (dorsal-root potentials, DRP) or from the ventral roots (ventral-root potentials, VRP).[2] These potentials spread electrotonically along the spinal cord and the dorsal and ventral roots. The DRP is a slow negative potential, preceded often by a positive deflection. There are no significant differences in these potentials evoked by stimulation of muscular or cutaneous nerves, in anesthetized animals or in nonanesthetized

[1] GASSER, H. S., and H. T. GRAHAM, *Am. J. Physiol.*, **103**, 303, 1933.

[2] BARRON, D. H., and B. H. C. MATTHEWS, *J. Physiol.*, **92**, 276, 1938; ECCLES, J. C., *Nature, London*, **153**, 432, 1944; **154**, 395, 1944; *J. Neurophysiol.*, **9**, 87, 1947; BREMER, F., and V. BONNET, *Arch. d. sc. physiol.*, **3**, 489, 1949; BROOKS, C. McC., *et al.*, *J. Neurophysiol.*, **13**, 157, 1950; BROOKS, C. McC., and M. G. F. FUORTES, *J. Physiol.*, **116**, 380, 1952.

decerebrated animals. The VRP is modified by anesthesia and differs according to whether it is evoked by stimulation of muscular or cutaneous nerves. It is made up of two groups of waves, both prominent in nonanesthetized animals. In an anesthetized animal a weak stimulus evokes a small slow rising negative potential, frequently preceded by a positive deflection. A stronger stimulus provokes the appearance of spikes superimposed on the slow potential. The spikes are synchronous when a muscular nerve has been stimulated; they are the electrical concomitant of a monosynaptic reflex.[1] When a cutaneous or a mixed nerve is stimulated, asynchronous spikes appear as a result of a plurisynaptic reflex. In most cases large positive potentials follow the propagated discharges. In nonanesthetized decerebrated animals there is no positive afterpotential, but a long-lasting negative potential related to the DRP. The early waves in the VRP are the electrical manifestation of reflex activity evoked by incoming impulses acting on the motor neurons. The slow potentials may be due to reverberating activity of interneuron circuits, or, as suggested by Barron and Matthews, to ionic changes external to the fibers produced at the axonic terminals following impulse activity. The process which gives rise to these slow potentials is undoubtedly of physiologic importance, because sustained and integrated reflex movements are observed only in conditions in which they appear. When they are absent, as in anesthetized animals, the excitability of the motor neuron is depressed after a short burst of activity (during the positive afterpotential) and only brief, rudimentary reflex movements can be performed.

[1] Renshaw, B., F. *Neurophysiol.*, **4**, 167, 1941; Lloyd, D. P. C., *J. Neurophysiol.*, **6**, 111 and 317, 1943.

Fig. 379. Electric potentials of the spinal cord evoked by stimulation of a dorsal root. (*Gasser, H. S., and H. T. Graham, Am. J. Physiol., vol. 103, p. 303, 1933.*)

CHEMICAL MEDIATORS OF SYNAPTIC TRANSMISSION

The existence of a chemical mediator (acetylcholine) in neuromuscular transmission of excitation and in the conduction of the nerve impulse has already been considered (Chaps. 67 and 68). Acetylcholine has also been found in the nerve centers, and it will be convenient to discuss at this point the part it plays in synaptic transmission of excitation, first in sympathetic ganglia and then in the central nervous system.

Chemical mediators in sympathetic ganglia Kibjakow[1] discovered the presence of acetylcholine in sympathetic ganglia and suggested that it played a part in synaptic transmission. Varying amounts have been reported; *e.g.*, 10 to 44 μg per gm. of fresh weight in the cat and 8 to 19 μg per gm. in the sheep. No ACh is found in a ganglion one or two weeks after its preganglionic fibers have been cut. Feldberg and Gaddum[2] perfused the superior cervical ganglion and showed that stimulation of preganglionic fibers liberated acetylcholine (Fig. 380). The preparation had to be eserinized in order to protect acetylcholine from the hydrolytic effects of

[1] Kibjakow, A. W., *Pflüger's Arch. f. d. ges. Physiol.* **232**, 432, 1933.

[2] Feldberg, W., and J. H. Gaddum, *J. Physiol.*, **81**, 305, 1934.

cholinesterase, which was also found in the ganglion. Antidromic stimulation of postganglionic fibers, or stimulation of fibers that did not end synaptically but passed through the ganglion, produced the release of only very small quantities of acetylcholine. Care must be taken, however, to maintain the preparation in good condition; otherwise, as it deteriorates, increasingly larger amounts are set free after preganglionic stimulation, and eventually this occurs even without stimulation.[1]

The output of ACh from a perfused eserinized ganglion submitted to repetitive preganglionic stimulation is at first about 100 μμg. per volley (superior cervical ganglion of cat) but then falls exponentially as stimulation is kept up. After 40 min. stimulation, the output of ACh is considerably reduced. If the ganglion is not eserinized, the output of ACh does not diminish with sustained stimulation; thus after a preliminary period of 40 min. preganglionic stimulation the output is the same as before stimulation. It seems that ACh is normally resynthesized from the products of its hydrolysis; eserine interferes with this process of synthesis, and the available stock of ACh is gradually depleted.[1]

Acetylcholine added to the perfusing fluid causes postganglionic, but not preganglionic, discharge.[2] Postganglionic discharge is also obtained by intravenous injection or by injecting small doses into the artery of the ganglion (*e.g.*, the carotid artery for the superior cervical ganglion). Larger doses provoke repetitive discharge, and very high doses have a depressing effect and may cause complete block.

Eserine and other anticholesterinases prolong the effect of ACh, potentiate subliminal doses of this drug and subliminal preganglionic stimulation, and cause repetitive discharge in response to a single preganglionic volley. Curare de-

Fig. 380. Release of acetylcholine provoked by preganglionic stimulation in the perfused superior cervical ganglion. *a*, above: perfused frog's heart (*Straub*); below: eserinized leech muscle; *A*, perfusion fluid collected during stimulation; *B*, 15μg acetylcholine; *C*, 30 μg acetylcholine; *D*, perfusion fluid collected without stimulating the ganglion. *b*, above: eserinized abdominal rectus muscle of frog; below: arterial blood pressure of cat; 1, perfusion fluid collected during preganglionic stimulation; 2, 50 μg acetylcholine; 3, 30 μg acetylcholine, 4, perfusion fluid collected without stimulating the ganglion. (*Feldberg, W., and J. H. Gaddum, J. Physiol., vol. 81, p. 305, 1934.*)

[1] Lorente de Nó, R., *Am. J. Physiol.*, **121**, 331, 1938.

[1] Perry, W. L. M., *J. Physiol.*, **119**, 439, 1953.

[2] Feldberg, W., and A. Vartiainen, *J. Physiol.*, **83**, 103, 1934; Bronk, D. W., S. S. Tower, and Y. Solandt, *Proc. Soc. Exper. Biol. & Med.*, **32**, 1569, 1935; Cannon, W. B., and A. Rosenblueth, *Am. J. Physiol.*, **119**, 221, 1937.

presses, and in relatively large doses completely blocks, ganglionic synaptic transmission; it does not, however, prevent the release of ACh by preganglionic stimulation, but antagonizes the effect of ACh, *i.e.*, it has the same activity as in the myoneural junction. ACh and anticholinesterases exert a decurarizing effect on the ganglion similar to that observed in muscle.

ACh depolarizes the ganglionic neurons. Normally this effect is of very short duration, because ACh is destroyed by cholinesterase. If it is protected by anticholinesterase or a large dose is injected, persistent depolarization causes ganglion block. ACh has the same effects as nicotine and certain quaternary ammonias (tetramethylammonium): the ganglionic neurons are first stimulated, then blocked. Curare alkaloids and other drugs such as pentamethonium, hexamethonium, and decamethonium block the ganglion without producing depolarization, perhaps by substrate competition with ACh.[1]

Potassium is released by preganglionic stimulation and by ACh; when added to the perfusion fluid, it releases ACh and causes postganglionic discharges. Adenosinetriphosphate (ATP) and creatinephosphate (CP) are powerful stimulants of the perfused ganglion. Their effects persist after denervation and degeneration of the preganglionic fibers, therefore they are not dependent on ACh release. The effects of ATP and CP are related to the phosphate in the molecule and are not suppressed by curare. The action of ATP and CP is similar to that of potassium.[2]

The adrenal medulla is in certain respects analogous with sympathetic ganglia (see Chap. 85), and its cells are innervated by preganglionic fibers of the splanchnic nerves. Stimulation of these nerves causes the release of acetylcholine, which can be collected in the adrenal vein, if the preparation is eserinized, *i.e.*, if destruction of acetylcholine by cholinesterase is prevented. Acetylcholine, moreover, provokes secretion of the adrenal medulla.

There is, therefore, good evidence that acetylcholine is released in sympathetic ganglia by preganglionic stimulation, and that it stimulates the ganglion cells.

[1] Paton, W. D. M., and W. L. M. Perry, *J. Physiol.*, **119**, 43, 1953.

[2] Feldberg, W., and C. Hebb, *J. Physiol.*, **107**, 210, 1948.

Chemical mediators in the central nervous system. Acetylcholine,[1] cholinesterase,[2] and choline acetylase[3] have been found in the tissue of the central nervous system. Acetylcholine concentration and enzymatic activity vary in different parts of the nervous system, but usually the three values are in agreement. Exceptions to this rule are the ventral roots which have high ACh and low cholinesterase, and the cerebellum, in which ACh is low and cholinesterase high. The dorsal roots and dorsal column of the spinal cord (f. gracilis and cuneatus) have low ACh and enzyme concentrations; the dorsal horns and the nuclei gracilis, cuneatus, and vestibularis have high values; in the internal capsule the values are low. Therefore the first neuron of the sensory pathway is noncholinergic, the second neuron is cholinergic, and the thalamic neurons are noncholinergic. The lower motor neuron is cholinergic, the pyramidal tract is noncholinergic, and the motor cortex has intermediate values. Acetylcholine and eserine stimulate the motor cortex; therefore some of its neurons, though probably not the Betz cells, are cholinergic. The ACh content of the brain increases during sleep or under the influence of anesthetics (ether) or narcotics (nembutal). Activity, such as is produced by emotional excitement, electrical stimulation, or drugs which provoke convulsions (metrazol, picrotoxin), lowers the ACh content of the brain. Normal values are recovered after a short period of rest; therefore, ACh is rapidly resynthesized.[4]

Acetylcholine applied locally increases the activity of the motor cortex,[5] and injected into the carotid in cats after section of the brain stem it evokes the awakening reaction in the electroencephalogram[6] (see Chap. 87). In the perfused spinal cord, eserine and prostigmine enhance the flexor reflex and depress the knee-jerk. Flexor responses to stimulation of the descending spinal

[1] Feldberg, W., *Physiol. Rev.*, **25**, 596, 1945.

[2] Nachmansohn, D., *Bull. Soc. chim. biol.*, **21**, 761, 1939; Burgen, A. S. V., and L. M. Chipman, *J. Physiol.*, **114**, 296, 1951.

[3] Feldberg, W., and M. Vogt, *J. Physiol.*, **107**, 372, 1948.

[4] Richter, D., and J. Crossland, *Am. J. Physiol.*, **159**, 247, 1949; Elliott, K. A. C., *et al.*, *Am. J. Physiol.*, **162**, 469, 1950.

[5] Sjöstrand, T., *J. Physiol.*, **90**, 41P, 1937.

[6] Bonnet, R., and F. Bremer, *Compt. rend. Soc. de biol.*, **126**, 1271, 1937; Bremer, F., *Compt. rend. Soc. de biol.*, **128**, 544, 1938.

tracts are usually facilitated by ACh and depressed by adrenaline, and extensor responses are facilitated by adrenaline and depressed by ACh.[1] Acetylcholine injected into the basilar artery evokes a burst of impulses from the ventral roots of the first cervical segment, and increases the area, reducing the latency of ipsilateral reflexes. The ventral root potential is increased in amplitude and rapidly reaches its peak. Eserine and prostigmine increase and prolong these effects of ACh.[2]

There is, therefore, sufficient evidence that the acetylcholine cycle functions in the central nervous system, that some of the neurons are cholinergic, and that ACh modifies synaptic transmission in the central nervous system, facilitating it in some cases, depressing it in others.

Sympathomimetic activity has been found in extracts of brain tissue.[3] This "sympathin" is a mixture of noradrenaline and adrenaline, the former predominating (93 per cent in cats and 86 per cent in dogs).[4] It has been found in all parts of the central nervous system, but unevenly distributed. The highest concentrations are in the hypothalamus (1.4 μg/gm. in the cat), midbrain (gray matter around the aqueduct), and medulla (floor of fourth ventricle and reticular formation); i.e., regions which contain centers of the sympathetic. Asphyxia and certain drugs which stimulate central sympathetic activity (ether, morphine, insulin) diminish the concentration of noradrenaline in the hypothalamus and midbrain. These facts are suggestive but the evidence is not sufficient to assign a definite functional significance to noradrenaline and adrenaline as chemical mediators of synaptic transmission in sympathetic centers.

SYNAPTIC TRANSMISSION

The nature of the processes taking place at the synapse, resulting in the transmission of a propagated impulse from the axon terminals of one neuron to the soma (dendrites and perikaryon) of another, or in the suppression of activity (inhibition) of the postsynaptic neuron, is one of the fundamental problems of neurophysiology. The processes of transmission appear to be essentially the same at myoneural junctions, synapses in autonomic ganglia, and central synapses, although there are features characteristic for transmission at each one of these sites.

There is sufficient evidence that, at myoneural junctions and ganglionic synapses, the impulse on arriving at the axon terminals releases acetylcholine (at some of the visceral myoneural junctions, adrenaline and noradrenaline are released). Acetylcholine acts on the postsynaptic membrane and causes depolarization, the electrical sign of which is the end-plate or synaptic potential. When depolarization covers a certain minimum area the synaptic potential reaches a critical level and a self-propagated excitatory process, the electrical sign of which is a spike potential, is initiated and spreads over the whole fiber or postsynaptic neuron. The membrane subjacent to several neighboring synaptic knobs must be activated simultaneously or within a very short interval for a spike to be fired off; otherwise there is only subliminal depolarization and a subliminal synaptic potential is observed. Acetylcholine probably acts on a receptor substance initiating changes in membrane permeability and ionic fluxes (inflow of Na^+, outflow of K^+) leading to depolarization and reversal of the membrane potential.

Evidence of chemical mediation of transmission at central synapses is not as satisfactory as it is for peripheral synapses. Eccles and his associates have demonstrated that in direct inhibition there is an increase in polarization in the motor-neuron membrane subjacent to the inhibitory synaptic knob. No electrical explanation of this actively produced anelectrotonus can be given; therefore, "inhibitory synaptic action is mediated by a specific transmitter substance that is liberated from the inhibitory synaptic knobs. . . . If inhibitory synaptic action is caused by a chemical transmitter, it seems probable that excitatory synaptic action in the central nervous system would also conform to the pattern of chemical transmission that seems to obtain with all other junctions."[1]

The nature and mode of action of the mediators are not known. The acetylcholine cycle has been demonstrated in the tissue of the central nervous system, and ACh can modify synaptic transmission in the central nervous system, but the part it plays in central synaptic transmission

[1] Bülbring, E., and J. H. Burn, *J. Physiol.* **107**, 289, 1948.

[2] *Ibid.*, **49**, 428, 1953.

[3] Holtz, P., *Acta Physiol. Scandinav.*, **20**, 354, 1950.

[4] Vogt, M., *J. Physiol.*, **123**, 451, 1954.

[1] Brock, L. G., J. S. Coombs, and J. C. Eccles, *Proc. Roy. Soc., London, s.B.*, **144**, 170, 1952; *J. Physiol.*, **117**, 431, 1952.

is not well understood. Several possibilities can be considered: (*a*) ACh released at different points of a neuron or at different neurons may exert different effects, *i.e.*, polarization at one place (inhibition) and depolarization at another (excitation); or (*b*) there may be two transmitters, one for excitation and one for inhibition, which may be released by the same cell, or more probably by different cells. There is no inherent impossibility for either of these hypotheses. Chemical mediators acting on myoneural junctions produce excitation at one place and inhibition at another, *e.g.*, acetylcholine stimulates the intestinal wall muscle and inhibits intestinal sphincters. The existence of more than one mediator in peripheral transmission is a well-known fact. Moreover, ACh and adrenaline influence central nervous activity in opposite ways,[1] and there is evidence that some of the central neurons are cholinergic and others are not. Adrenergic nerve endings release adrenaline and noradrenaline; it is possible, therefore, that the same neuron may release two chemically related mediators with different activity or that a mediator may be slightly modified after release, and its activity thus changed.

Bibliography will be found at the end of the next chapter.

[1] Bülbring, E., J. H. Burn, and C. R. Skoglund, *J. Physiol.*, **107**, 289, 1948.

CHAPTER 70

Reflex Coordination

A SIMPLE, ISOLATED REFLEX is a theoretical abstraction; it is convenient in the analysis of the functions of the nervous system, but it has no real existence. A particular reflex is conditioned by the activity of other nerve centers, and activity in a given reflex arc conditions that of others. This reciprocal action of reflexes on each other is possible because there are synaptic connections between all the paths, as is demonstrated by the effects of strychnine and tetanus toxin. These substances destroy the central inhibitory state, and stimulation of an afferent nerve no longer produces contraction of certain muscles and relaxation of their antagonists, but only contraction. In an advanced stage of intoxication it is possible to provoke convulsive contraction of all the muscles in the body by stimulation of any receptive field. In normal conditions only reflexes provoked by very weak stimuli or in widely separated areas exert no influence on each other. Irradiation of the effects obtained by stimulating an afferent path is considerably reduced by the action of certain drugs, such as chloroform and chloral, and during depression following section of the spinal cord (see "Spinal shock," Chap. 81).

The central nervous system normally functions as an integrated unit, and the combination of reflexes, taking place simultaneously or successively, into orderly responses is a fundamental aspect of nervous coordination. Integration of nervous activity is carried out at different levels of increasing complexity. The first and simplest integration is performed in the spinal cord. It can be examined by suppressing the influence of higher centers. For this purpose a spinal animal is used, *i.e.*, an animal that has been decapitated by section below the medulla oblongata. Artificial respiration must be performed, because the respiratory center is situated above the section. A "chronic" spinal animal can be kept alive with adequate care for several months if the section is made in the lower cervical region below the emergence of the roots forming the phrenic nerves.

The principle of reciprocal innervation. A fundamental process of reflex coordination is made evident on examination of the flexion reflex. This reflex consists in flexion of a limb in which a cutaneous nerve is stimulated centripetally; at the same time the limb on the other side is extended (crossed-extension reflex). If the activity of the different muscles that take part in the reflexes is recorded, the flexors of the ipsilateral limb will be seen to contract, and the extensors to relax (Fig. 381); the flexors on the contralateral limb relax and the extensors contract. Stimulation of the receptive field awakens activity in certain motor units (the protagonists of the movement) and inhibits others (the antagonists) (Fig. 382). There is close correlation in the intensity of the opposite processes. Thus, if the flexors are 100 per cent stimulated, the extensors are 100 per cent (*i.e.*, completely) relaxed; if the flexors are 25 per cent stimulated, activity in the extensors is reduced by 25 per cent. Smooth performance of a movement is the result of interaction between excitation and inhibition; one process is as important as the other. The part played by excitation is obvious; that of inhibition is demonstrated in the following experiment: All the nerves to the eye muscles, except the sixth nerve, which innervates the external rectus muscle, are cut on one side, *e.g.*, the left. The eye at rest is turned outward by tonic contraction of the external rectus. The animal's attention is drawn to an object (food or light) placed in the right eye field. The right eye is accurately fixed on the object, and the left eye is brought to the mid-line, *i.e.*, looks straight

a

b

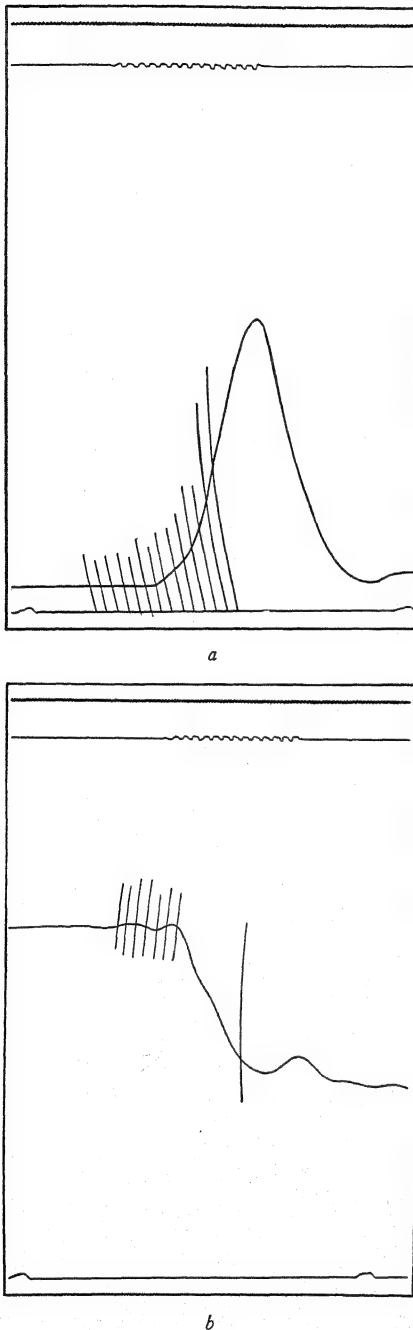



FIG. 381. Reciprocal innervation of antagonists. Flexion reflex evoked by stimulation of the ipsilateral internal saphenous nerve by a series of induction shocks. *a*, contraction (excitation) of the flexor muscle of the knee. *b*, relaxation (inhibition) of the extensor muscle of the knee. (*Sherrington, C. S., "The Integrative Action of the Nervous System," Yale University Press, New Haven, 1926.*)

forward. If the object is then placed in the left eye field both eyes are fixed on it, and will follow it as it is slowly displaced toward the right. When the object passes the mid-line, the right eye continues to follow it, but the left eye remains motionless, looking straight forward. The movements of the left eye toward the right are due exclusively to inhibition of the postural tonus of the external rectus, because all the other muscles have been denervated. Moreover inhibition is accurately coordinated with excitation of the right external rectus and inhibition of the right internal rectus, until the external rectus is fully relaxed and the eyeball adopts the resting position with the pupil directed straight forward. Whenever an integrated phasic movement is performed, there is a similar coordination of excitation in the protagonists and inhibition of the antagonists. Sherrington has called this the "principle of reciprocal innervation of antagonistic muscles."

In coordination of movements that involve muscles on both sides of the body, in some cases the principle of reciprocal innervation extends to symmetrical muscles. Thus in the flexion and crossed-extension reflexes evoked by stimulation of a cutaneous nerve of a limb, the ipsilateral flexors contract and the contralateral flexors relax, although these muscles are not, strictly speaking, antagonists. In other cases, *e.g.*, in symmetrical movements such as the clasping reflex of the toad and galloping movements in mammals, coordination takes place by *identical innervation*, *i.e.*, symmetrical pairs of muscles are excited or inhibited together. Muscles that are antagonists in certain movements act as synergists in others. For example, the flexors and extensors of a limb, which are antagonists in flexion and crossed-extension reflexes, contract together in the postural reflex which fixes the limb into a supporting column (see "Postural reflexes," Chap. 81). In this case also they are coordinated by the process of identical innervation.

The role of the receptors. The receptors, placed on the afferent end of the reflex arc, act as the first "analyzers" of changes in the external or internal environment.

This is due to their "specific irritability" (see Chap. 71), *i.e.*, each receptor has a low threshold for one kind of stimulus and a high threshold for others. Moreover, each receptor is more closely connected, by reason of the physiologic archi-

tecture of the nervous system, with certain motor mechanisms than with others. For example, a painful stimulus applied on the plantar surface stimulates one type of receptor and provokes the defensive flexor reflex; pressure on the same surface stimulates another type of receptor and the response is an extensor thrust.

Another factor which may condition the response is the site on the receptive surface on which the stimulus is applied. Thus, in the spinal cat stimulation of the skin of a limb evokes flexor responses except when the area over the antagonistic extensors is stimulated. Moreover, stimulation of the skin over a flexor muscle inhibits it.[1]

The "receptive field" of a reflex is the aggregate of receptors that respond to the same type of stimulus, the application of which produces a series of reflexes very similar, but not quite identical, to each other. In each field there is a central focal area where the threshold is lowest; in the periphery of the field, stronger stimuli must be applied to obtain a response. For example, the receptive field of the flexion reflex in the dog extends over the surface of the whole limb, but its focal area lies in the toe pads.

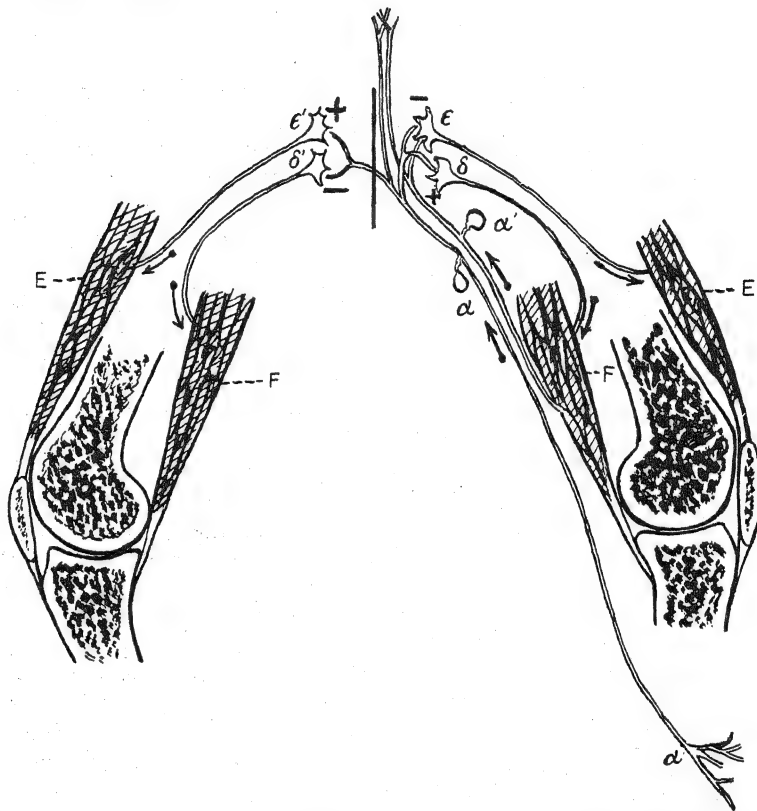



FIG. 382. Diagram of reciprocal innervation of antagonists in the flexion and crossed-extension reflexes. Stimulation of the afferent fiber α evokes ipsilateral excitation (+) in the flexors (*F*) and inhibition (−) in the extensors (*E*) (flexion reflex), and contralateral excitation (+) in the extensors and inhibition (−) in the flexors (crossed-extension reflex). (*Sherrington, C. S., "The Integrative Action of the Nervous System," Yale University Press, New Haven, 1926.*)

The threshold of a reflex can change from one moment to another, according to the physiologic condition of the animal. The change takes place in the whole field, so that the relative excitability of its different areas remains unchanged, but this relative value may be upset by "local fatigue" at a point that has been submitted to prolonged stimulation.

The receptive field of a reflex does not coincide with that of a dorsal root. Usually the afferent fibers enter the spinal cord by several roots; field and reflex are, therefore, plurisegmental in

[1] HAGBARTH, K. E., Excitatory Skin Areas for Flexor and Extensor Motoneurons, *Acta physiol. Scandinav.*, **26**, Suppl. 94, 1952.

extension. Conversely, in the same dorsal root there are afferent fibers of many superficial and deep receptive fields.

The local sign of reflexes. Reflexes are responses, the purpose of which is to adapt the organism to changes in the environment. The "purpose" of complex reflex patterns is usually clearly evident; *e.g.*, the flexion and crossed-extension reflexes evoked by a painful stimulus are directed to the removal of the limb from the harmful agent. The "purpose" of simple reflexes, however, is not always easy to understand. Adequacy of adaptation is clearly seen in the so-called "local sign" of reflexes, which consists in the orientation of the reflex movement in relation to the position of the stimulus. For example, the scratch reflex is made up of two components: (*a*) rhythmic alternating flexion and extension of one of the hind limbs; (*b*) sustained posture, the trunk being sent toward the side on which the stimulus is applied, the head slightly extended and rotated, both forelimbs and the hind limb of the opposite side extended, and the hind limb on the stimulated side brought more or less forward, according to the situation of the place stimulated, this being the local sign. Here, adequacy of response (scratching to remove the stimulus) is evident.

"In the execution of spinal reflexes the most important afferent factor as regards local sign is the afferent channel from the place of initiation of the reflex."[1] The correct direction of the reflex demands a state of equilibrium in the central nervous system, which is lost in spinal shock; in this condition there are gross errors in localization. Deafferentation also disturbs adequate orientation of reflex movements, even though the response can be obtained with facility equal to or greater than in the normal animal. In this condition afferent impulses from the activated muscles that serve to guide the movement are missing.

SIMULTANEOUS COMBINATION OF REFLEXES

Many types of receptors are found close to each other in all parts of the body. In the skin there are tactile, temperature, and pain receptors, and in the deeper structures there are the different kinds of proprioceptive receptors. The stimulus may act on a single receptor or on several receptors of the same or different types. When several receptors are stimulated and the afferent impulses act in the same sense, the effect of each one is reinforced by that of the others. Thus, a substance that stimulates the taste buds in the tongue produces reflex secretion of saliva; if this substance dries the mouth, this fact provides a new stimulus, which also provokes secretion of saliva; the reflexes reinforce each other. When impulses from different receptors use the same final common path in the same direction, either for excitation or for inhibition, their reflex effects are summated and they are called "allied reflexes."

The reflex arcs of a type reflex form an aggregate of allied reflexes. Impulses arriving at the center from each one of the receptors excite a certain number of motor neurons, and around them other neurons are excited subliminally. The subliminal fringe may be of low intensity and widely spread (*e.g.*, in labyrinthine reflexes) or narrow and of relatively high intensity (*e.g.*, in the flexion reflex). Impulses that arrive from other receptors to this subliminal fringe are facilitated. Thus two subliminal stimuli applied simultaneously, or with a short interval, to two different receptors can summate and a response will be evoked. Sherrington has called this type of facilitation "immediate spinal induction." Induction is more marked when the central connections of the receptors are close to each other.

When impulses from two receptors converge to a motor center and produce opposite effects, *i.e.*, excitation and inhibition respectively, the reflexes are known as "antagonistic." The response in each case is the result of the algebraic sum of the opposite central states of excitation and inhibition. For example, if a limb that is in extension as the result of a crossed-extension reflex is stimulated with a harmful stimulus, the state of excitation in the motor centers of the extensors is suppressed and the flexor muscles are stimulated (flexion reflex). If the stimulus is weak, there will be no apparent flexion, but the strength of contraction in the extensors will be reduced; if the stimulus is strong, the flexion reflex will be preponderant and the limb will be flexed. In both cases there is an occult reflex covered by the apparent effect of the stronger reflex. Inversion of the response is provoked by

[1] SHERRINGTON, C. S., "The Integrative Action of the Nervous System," Yale University Press, New Haven, 1926, p. 251.

a change in the relative strength of the stimuli, which may be due to a change of posture, to a third reflex converging on the same center which reinforced either the c.e.s. or the c.i.s., or to other circumstances. Experimental stimulation of an afferent nerve artificially increases occult reflexes, because many afferent paths producing opposite effects are stimulated simultaneously. In normal conditions the stimulus falls on a receptive surface and, because of specific irritability, excites some receptors and not others.

Graduation of the strength of reflex contraction. Stimulation of an afferent nerve of a limb produces almost simultaneous activation of all the motor centers of the flexion reflex, which thus takes place suddenly (jet or *d'emblée* reflexes). This is due to simultaneous stimulation of a great number of afferent paths. When, as normally occurs, the reflex is initiated by stimulation of receptors, the reflex develops gradually; there is progressive "recruitment" of motor units.

The strength of the response is graded by two mechanisms:

1. A variable number of motor units takes part in the reflex. At the beginning only two or three may be activated; later the number increases progressively.
2. An increase in the frequency of the impulses discharged by the center increases the tension developed by each motor unit. For example, in the phrenic nerve 15 impulses per second have been registered in shallow breathing and 60 to 80 impulses per second in deep breathing.

Both mechanisms are closely associated and the response can thus vary in extension and intensity.

The motor center is a mechanism for summation of c.e.s. and c.i.s. produced by afferent impulses that converge to it. Some neurons may be stimulated maximally and discharge with a frequency sufficient to provoke the strongest tetanic contraction possible. Other neurons, close to the former, are also stimulated above the threshold, but less intensely, and these discharge at a lower frequency, producing in the corresponding muscle fibers a more or less incomplete tetanus (subtetanic contraction). Neurons that are farther away from the focus are stimulated subliminally, and even more distant neurons remain quiescent (Fig. 369).[1] The arrival of further impulses through either the same or other paths intensifies and extends the c.e.s. The effect on maximally stimulated neurons cannot be any greater, because they are "occluded," but submaximally stimulated neurons discharge at a higher frequency, and those of the subliminal fringe also discharge. This ascending scale of excitation takes place at the beginning of a reflex until the latter is fully developed; as it wanes, the process is reversed and motor units drop out progressively until the reflex ceases completely.

IRRADIATION OF REFLEXES

Certain reflexes take place within a small segment of the spinal cord; the afferent impulses enter through one or a few dorsal roots, and the motor centers are situated in one or a few spinal segments. These are called *short reflexes*, and they develop following a fairly constant pattern.

Other reflexes, *e.g.*, the scratch reflex, have arcs that occupy many segments of the spinal cord. They are *long reflexes*, and their development may show considerable differences in individual instances.

Sherrington's laws of reflex irradiation.[2] The spread of a reflex to an increasing number of motor units takes place according to a certain pattern, which is described by Sherrington's five laws of reflex irradiation.

1. The degree of reflex spinal intimacy between afferent and efferent spinal roots varies directly as their segmental proximity. Motor effects are obtained more easily through the efferent roots of the stimulated and neighboring segments than through the roots of more distal segments (law of spatial proximity).
2. For each afferent root there exists in its own segment a reflex motor path of as low a threshold and of as high a potency as any open to it anywhere else. It must be kept in mind that afferent fibers make central connections in several segments with many motor neurons, either directly or through internuncial neurons.

[1] Probably in all neurons of the central nervous system there is always some degree of excitation or inhibition produced by "bombardment" from internuncial neurons. It is perhaps not strictly accurate therefore to speak of neurons in a state of complete rest or quiescence in a normal nerve center.

[2] SHERRINGTON, *op. cit.*, p. 158.

3. The motor mechanisms of a segment are unequally accessible to the local afferent channels. As judged by excitatory effects, synaptic resistance is not the same in all the paths, but if inhibitory effects are taken into account, the reflex connection is found to be wider.
4. The motor neurons simultaneously discharged by a spinal reflex innervate synergic and not antergic muscles.
5. The spinal reflex movement elicited in and from any one spinal region will exhibit much uniformity despite considerable variety of the locus of incidence of the exciting stimulus. Even when the receptive field is spread over several spinal segments, the motor response is fundamentally the same.

Reflex figures. (Fig. 383.) Long reflexes spread following a path determined by laws of spatial proximity and synaptic resistance. Thus certain "reflex figures" are established, which can be observed in the "spinal animal" (*i.e.*, after section of the cord below the medulla) or the "decerebrate animal" (*i.e.*, after section of the brain stem between the colliculi).

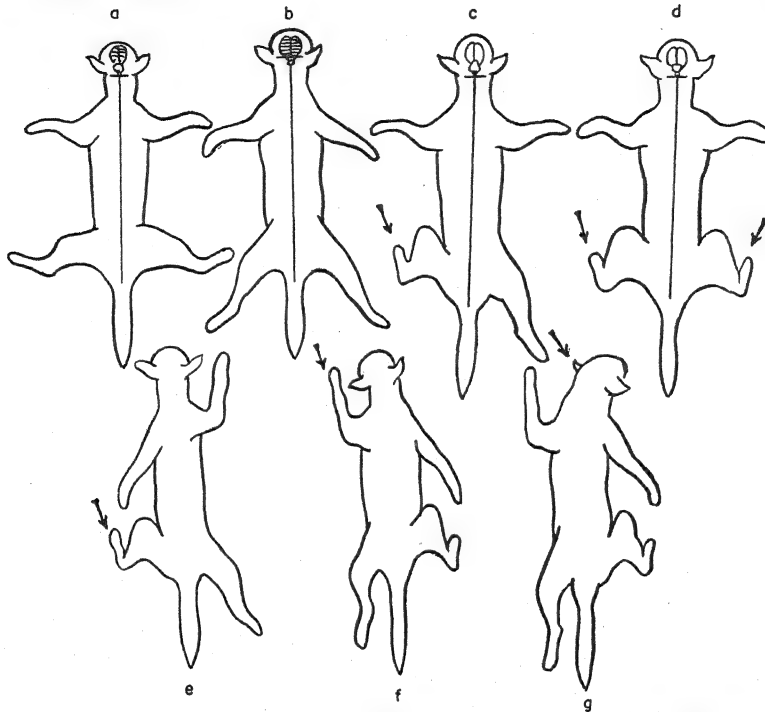



FIG. 383. Reflex figures. *a*, position of spinal animal; *b*, position of animal in decerebrate rigidity; *c*, change of attitude in spinal animal produced by a stimulus on the left hind foot; ipsilateral flexion and crossed-extension reflexes; *d*, simultaneous stimulation of both hind feet; bilateral flexion reflexes inhibiting the crossed-extension reflexes; *e*, stimulation of left hind foot irradiated to forelimbs, extension of the ipsilateral and flexion of the contralateral limb; *f*, stimulation of the left forefoot irradiated to the neck and hind limb, ipsilateral extension and contralateral flexion; *g*, stimulation of the left pinna irradiated to the neck and limbs. (*Sherrington, C. S., "The Integrative Action of the Nervous System," Yale University Press, New Haven, 1926.*)

In the spinal animal, on increasing the strength of the stimulus of a flexion reflex started in a hind limb, the following sequence is observed: (*a*) extension of the contralateral hind limb; (*b*) extension at the elbow, and retraction at the shoulder, of the ipsilateral forelimb; (*c*) flexion at the elbow, extension at the wrist, and a certain degree of protraction at the shoulder, of the contralateral fore limb; (*d*) twisting of the neck toward the stimulated side, with snapping of the jaws and switching of the tail. The animal appears to try to avoid proximity to the stimulus, first by a flexion reflex, which reduced to its minimum consists in the contraction of a few motor units flexing the knee; then by movements of locomotion; and

finally by trying to remove the nocuous agent with its teeth. From the other limbs and the ears similar reflex figures may be evoked.

In the mammal it is easier to obtain irradiation along the spinal cord in a caudal than in a cephalic direction. Movements of the limbs and tail are frequently obtained by stimulating the pinna, but the reverse rarely occurs. Sometimes in the course of irradiation one or more segments are skipped over; from the head the movement spreads to the hind limb without involving the forelimb, or the tail switches while the limbs are still. Reflexes, however, can also spread in the opposite direction, *e.g.*, ear twitching on caudal stimulation.

Crossing over to the opposite side of the spinal cord takes place more easily at certain levels than at others; this varies with the species and with the condition of the animal. Thus, irradiation to the opposite side is more frequent at the level of the hind limb than at that of the forelimb in the dog and the cat. In the frog and toad, spreading from one forelimb to another is facilitated in spring and summer, when the sexual reflex is active. This reflex consists in bilateral flexion of the forelimbs on stimulation of the skin on their internal aspect or on the chest. This response is symmetrical; in other instances irradiation produces an asymmetrical response, *e.g.*, the crossed extension which accompanies the flexion reflex.

Irradiation does not usually take place gradually or smoothly, but by "jumps," because the different allied arcs have approximately the same threshold; therefore when the stimulus is sufficiently strong to activate one of them it soon reaches a level that activates the others. A considerably greater increase in the strength of the stimulus is necessary to excite centers of other reflexes with a higher threshold.

Lloyd[1] has determined the paths and internuncial neurons that take part in long reflexes. This was done by recording the action potentials of spinal nuclei and fiber paths, using extremely fine electrodes inserted in the spinal cord at different levels. Unilateral reflexes (*e.g.*, the scratch reflex) are mediated by the fasciculus proprius in the ventral columns of both sides and the ipsilateral lateral column. Bilateral reflexes (*e.g.*, hand-foot reactions) are mediated by propriospinal fibers on both sides of the cord. Internuncial neurons of these reflexes are grouped in the ventral horn, near the motor neuron (nucleus proprius cornu anterioris). Collaterals of fibers in the ventral column end on the nucleus of the anterior commissure; axons emerging from this nucleus pass through the anterior commissure and converge on the motor neurons of the opposite side. The fasciculus proprius extends into the brain, under the name of "medial longitudinal fasciculus." It is of great importance in reflexes mediated by the cranial nerves.

[1] Lloyd, D. P. C., *Physiol. Rev.*, **24**, 1, 1944.

SUCCESSIVE COMBINATION OF REFLEXES

Reflexes and reflex figures follow each other in orderly sequence. There is no fusion of responses; one reaction replaces another without confusion. This is evident in alternation of antagonistic reflexes. For instance, flexion takes the place of extension because the predominance of c.e.s. and c.i.s. is inverted in the centers. Coordination also occurs when one reflex replaces another with which it is in part allied. For example, when a scratch reflex provoked by stimulation of the anterior part of the receptive field is replaced by a scratch reflex provoked by a stronger stimulus in a posterior part of the field, the rhythmic alternating movements of flexion and extension are not altered, but the postural contraction is changed so that the leg is directed to the spot where the second stimulus is applied, and not to an intermediate spot between the two stimuli.

Factors that condition predominance of a reflex. A predominating reflex may give place to another in certain conditions. The following factors play a part in bringing about a change in the reflex pattern: (*a*) successive spinal induction; (*b*) fatigue; (*c*) relative intensity of different stimuli; (*d*) type of reflex; (*e*) plasticity.

Successive sbinal induction. After a final common path has been used in one sense, the response of a reflex that uses it in the opposite sense is augmented. For example, a crossed-extension reflex is repeatedly provoked at certain intervals. Between two stimuli, another stimulus is applied to provoke a flexion reflex, which has the opposite effect on the motor centers. Crossed extension following flexion is augmented, although the strength of the stimulus has not changed. This augmented capacity of response lasts for some time and disappears gradually. It is not an effect of rest, because it does not occur

Time in secs.

Signal

a

Time in secs.

Signal

b

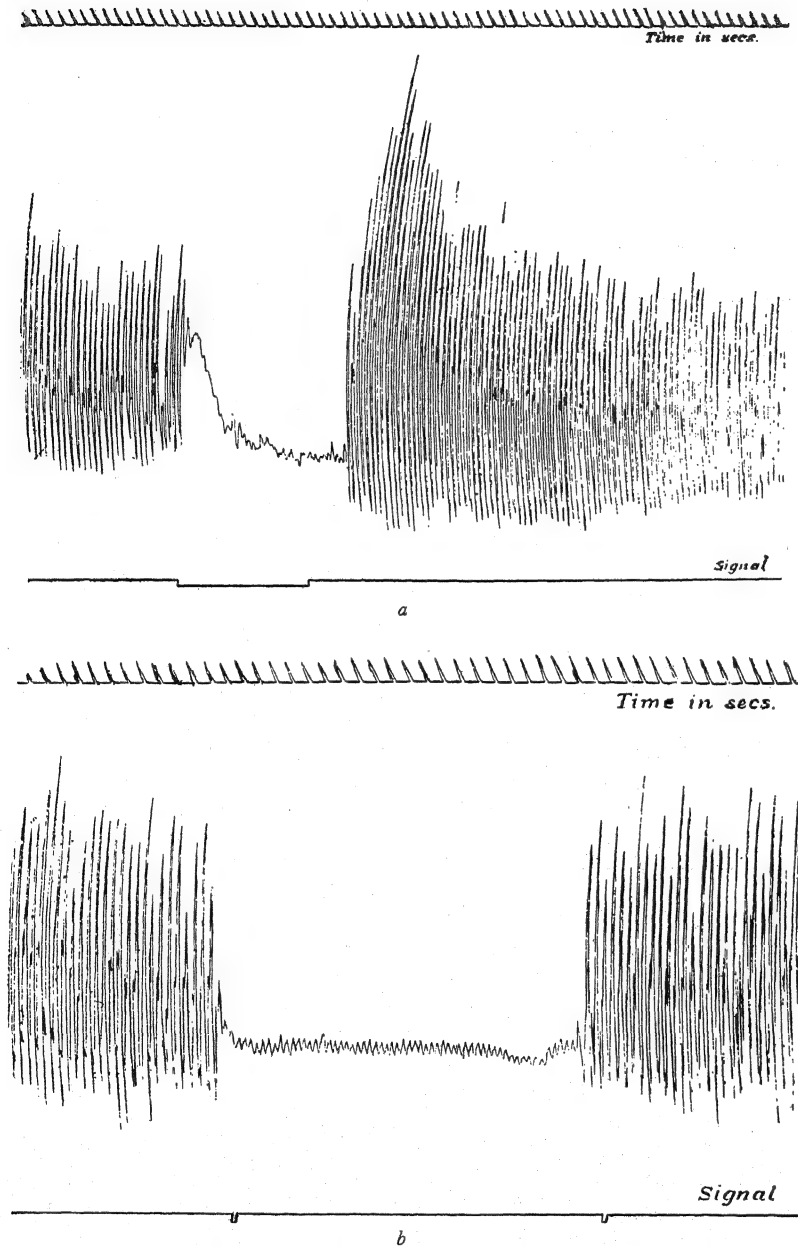



Fig. 384. Successive spinal induction. *a*, "mark-time" reflex (upstrokes correspond to flexion and downstrokes to extension) inhibited by stimulation of the tail; when the inhibitory stimulus ceases the reflex reappears with increased amplitude and quickening of the movements (successive induction). *b*, the same reflex interrupted by removing the stimulus between the two marks; when the stimulus once more acts, the reflex response reappears with the same strength as before (there is no induction). (*Sherrington, C. S., "The Integrative Action of the Nervous System," Yale University Press, New Haven, 1926.*)

when the reflex is not inhibited but merely interrupted by ceasing to apply the stimulus (Fig. 384). The greater response is due to a stronger c.e.s. This mechanism, which has been called "successive spinal induction" or "indirect spinal induction" by Sherrington, facilitates the use of the common path by different reflexes and prevents one afferent path from appropriating a motor center permanently.

Rebound is similar to successive spinal induction. Withdrawal of a stimulus or a sudden reduction in its strength may be followed not by a mere lapse of the response but by an outbreak of fresh activity, which may be in the same direction as before but more considerable, or may be in the opposite direction; *e.g.*, strong contraction follows moderate contraction, or marked relaxation follows less marked relaxation, or contraction replaces inhibition. Rebound has a relatively long latent period, lasts a long time, begins and ends gradually, and is easily inhibited. It is peculiar to the motor centers and is independent of the higher centers, since it is observed in the spinal animal. It is not due to stimuli coming from the active muscles, because it can be observed after deafferentation, nor is a change of tension or length in the muscles a necessary factor. It is marked when the reflex is provoked by stimulating an afferent nerve instead of the corresponding receptor, because excitatory and inhibitory fibers are activated simultaneously and occult reflexes are provoked. On altering the strength of the stimulus, the preponderance of a reflex may vary and the final response is thus changed. Reflexes provoked by normal stimulation of the receptors also show rebound. Thus pressure on the palate and the upper dental arch opens the mouth, partly by contraction of the digastric muscle but principally by inhibition of the temporal and masseter muscles. On withdrawing or diminishing the pressure, the mouth is quickly closed by a sharp contraction of the muscles that raise the jaw. This mechanism plays a part in mastication (see Chap. 36).

Successive spinal induction is important in alternating and rhythmic reflexes and in reflexes that serve to compensate for the effects of a previous reflex.

Fatigue. When an afferent path has had the use of a final common path for some time, it gradually loses its control by "fatigue," in spite of the fact that the motor unit can fully respond to stimulation through another afferent path. The threshold of the reflex increases gradually, and the stimulus becomes subliminal. Local fatigue disappears after a few minutes' rest.

Relative intensity of different stimuli. When several stimuli act simultaneously, the stronger stimulus takes possession of the final common path, but its efficiency gradually diminishes because of local fatigue in the centers and because a receptor discharges impulses at a lower frequency as time passes, in spite of the continuous action of the stimulus (see "Adaptation," Chap. 71). Successive spinal induction also contributes to increase the relative efficiency of a stimulus with respect to another stimulus that has been evoking a response.

Type of reflex. Reflexes that are accompanied by an intense sensation of pain or pleasure ("affective tone") predominate over others of lower affective tone. Thus phasic reflexes produced by nocuous stimuli easily replace postural reflexes. Predominance of this nature is of short duration, because this type of reflex is soon fatigued, and the receptors show marked adaptation. Postural reflexes are easily inhibited, but they can hold a motor path for a long time because they are not easily fatigued, and the corresponding receptors have a moderate degree of adaptation.

Plasticity, *i.e.*, changes in the capacity to transmit impulses caused by previous "experience" (use or disuse) of a synaptic path. Eccles and McIntyre[1] have given a good demonstration of this in monosynaptic reflexes. Two dorsal roots (*e.g.*, the seventh lumbar and the first sacral, on the left side) were cut distal to the ganglion, taking care not to damage the ganglion or the ventral root. Forty days later, *i.e.*, after a long period of disuse of the synapses of the cut roots, these were stimulated and the reflex response was compared with that obtained by stimulation of the corresponding normal roots on the opposite side. Reflex discharges in the ventral roots or motor nerves of the operated side were much smaller than those obtained from the normal side. Posttetanic potentiation gave a smaller enhancement on the operated side, but it had a more prolonged effect. The response from the root immediately rostral to the cut one was larger than from the corresponding root on the intact side, as if some compensatory reaction to the operative disability had developed.

There are many mechanisms for replacing

[1] Eccles, J. C., and A. K. McIntyre, *Nature, London*, **167**, 466, 1951; *J. Physiol.*, **121**, 492, 1953.

one reflex figure by another. When several stimuli are applied at the same time, factors that change preponderance come into play. When they are applied successively, if they produce allied reflexes, immediate spinal induction facilitates the effect of successive stimuli; if they produce antagonistic reflexes, successive spinal induction comes into play.

In certain instances the effect of one reflex acts as a stimulus for another. In swallowing, for example, a reflex sends the bolus into the esophagus and distends it. This distention produces contraction of the circular muscle of the esophagus above the bolus and relaxation below. The bolus therefore progresses, distends a lower segment, and stimulates another reflex.

The possibility of changing the use of a final common path from one movement pattern to another, or from excitation to inhibition, is essential for adaptation to the environment by adequate response to changes that may occur in the environment.

BIBLIOGRAPHY

Blair, H. A., Conduction and Synaptic Transmission in the Nervous System, *Ann. Rev. Physiol.*, **12**, 399, 1950.

Brooks, C. McC., and M. G. F. Fuortes, Excitation, Conduction and Synaptic Transmission in the Nervous System, *Ann. Rev. Physiol.*, **14**, 363, 1952.

Bullock, T. H., Conduction and Transmission of Nerve Impulses, *Ann. Rev. Physiol.*, **13**, 261, 1951.

Creed, R. S., D. Denny-Brown, J. C. Eccles, E. G. T Liddell, and C. S. Sherrington, "Reflex Activity of the Spinal Cord," Oxford, New York, 1939.

Feldberg, W., Present Views on the Mode of Action of Acetylcholine in the Central Nervous System, *Physiol. Rev.*, **25**, 596, 1945.

Feldberg, W., The Role of Acetylcholine in the Central Nervous System, *Brit. M. Bull.*, **6**, 312, 1950.

Fulton, J. F., "Physiology of the Nervous System," 3d ed., Oxford, New York, 1949.

Larrabee, M. G., and C. Edwards, Excitation and Conduction in the Nervous System, *Ann. Rev. Physiol.*, **15**, 283, 1953.

Lloyd, D. P. C., Functional Organization of the Spinal Cord, *Physiol. Rev.*, **24**, 1, 1944.

Royal Society, London, A Discussion on Excitation and Inhibition, *Proc. Roy. Soc., London, s.B.*, **140**, 169, 1952.

Sherrington, C. S., "The Integrative Action of the Nervous System," Yale University Press, New Haven, 1926.

Transmission synaptique ganglionnaire et centrale, *Archives internat. de physiol.*, **59**, 475, 1951.

CHAPTER 71

The Mechanism of Sensation

THE ORGANISM COMES into contact with its environment through receptors distributed on its surface; stimulation of these receptors is followed by reactions of adaptation to the outside world. In complex organisms there are also many receptors in the depth of the tissues, which are stimulated by changes within the body. Stimulation of this second type of receptor is followed by reactions that integrate the organism into a coordinated unit. Excitation of the receptors is the initial step in sensation, which is the raw material used by the mind to build up its knowledge of the universe around it and of its own inside world.

Knowledge of the material universe is conditioned—therefore limited—by the competence of the receptors. Man has used his intelligence to increase the scope of his receptors, in order to acquire knowledge of phenomena that do not stimulate them directly. This is done by converting energy set free by these phenomena into other forms of energy that act as stimuli. Thus rays of shorter or longer wavelength than light (less than 330 and more than 900 mμ) do not excite the human photoreceptor; nevertheless accurate knowledge of ultraviolet rays, and of those of even shorter wavelength such as x-rays and cosmic rays, has been obtained. In a similar way infrared rays and rays of longer wavelength, such as those used in broadcasting, have been studied. The natural capacity to differentiate quality and quantity in the stimuli has been considerably increased; thus in optimal conditions the human skin cannot appreciate differences in temperature of less than 0.2°C., while thermoelectric piles can register changes in temperature of 0.0000001°C.

Insufficiency of one sensory apparatus can be compensated by training another. Compensation takes place subconsciously and is remarkably efficient in some cases. Thus in locomotor ataxia, a disease in which deep sensations are lost, the patients can keep their balance as long as they see, because visual sensations compensate for the loss of those from the muscles and tendons, but on shutting their eyes they fall down. Persons who are born deaf (deaf-mutes) learn to speak when they are taught, through visual and tactile senses, how to use their muscles in the complicated movements necessary to articulate words. The blind learn to read the Braille system, in which tactile sensations replace visual ones. Helen Keller, when 19 months old, lost her sight and hearing owing to disease, yet she not only learned to read and write, but was able to acquire a vast amount of knowledge and cultivate her exceptional intellect, mainly on the basis of tactile sensations.

CLASSIFICATION OF THE RECEPTORS

Sherrington has distinguished three divisions in the distribution of the receptors. The deep-lying ones are in what he calls "the *proprioceptive field*, because its stimuli are, properly speaking, events in the microcosm itself, and because that circumstance has important bearing upon the service of its receptors to the organism."[1] The receptors distributed on the surface can be divided into those of the outer surface, the *exteroceptive field*, and those of the inner, or alimentary and respiratory, surface, the *interoceptive field*. This classification has been slightly modified. Proprioceptors, stimulated by pressure, tension, and stretching, together with the exteroceptors, form part of the somatic division of the nervous system. Interoceptors

[1] SHERRINGTON, C. S., "The Integrative Action of the Nervous System," Yale University Press, New Haven, 1926, p. 317.

of the inner surface are visceral receptors, similar to those found in the heart, blood vessels, bladder, and other viscera. These visceroceptors are stimulated by changes occurring in the organism and, with the proprioceptors, give information of internal changes. The different receptive fields can be summarized as follows:

Somatic afferent system { Exteroceptive / Proprioceptive } / Visceroceptive } Interoceptive

Exteroceptors. The exteroceptive field is the richest in receptors, some of which are highly differentiated. Exteroceptors give rise to clearly defined, well-localized sensations, and provoke reactions that usually are conscious. Two groups can be recognized in this field, the *distance receptors* or *telereceptors* and the *contact receptors.*

Distance receptors. Stimulation of this type of receptor gives rise to sensations that are localized, or psychically "projected," outside the organism and referred to the origin of the stimulus. Sight and hearing form part of this group.[1] These receptors have the following characteristics:

1. They are placed in the head, *i.e.*, the highest and most forward part of the body; therefore they come into contact with a great number of stimuli.
2. They are served by a much larger number of afferent fibers than other receptors. Thus the photoreceptor is innervated by approximately a million fibers in the optic nerve, and these form part of secondary neurons, each one of which is connected with several primary retinal neurons. On the other hand all the exteroceptive, interoceptive, and proprioceptive fields of the head, trunk, and limbs are served by little more than half a million afferent fibers, *i.e.*, those of the dorsal spinal roots and the cranial nerves not attached to the distance receptors.
3. They have an extensive representation in the cerebral cortex and appear to be particularly important in the development of the nervous system. According to Sherrington, "the brain is always the part of the nervous system which is constructed upon and evolved upon the distance receptors."[1]
4. They are widely connected with the different parts of the nervous system by many internuncial neurons.
5. Distance receptors evoke reactions that can be complex and widespread, sometimes involving the whole organism. These reactions are of anticipatory nature, preparatory to other final reactions that complete adaptation.
6. The "affective tone," *i.e.*, feeling of pleasure or pain, accompanying sensations evoked from these receptors is usually low, but it may be considerable when they are associated with previous experience of high affective tone.

[1] The receptors of smell and taste are placed on the internal surface, therefore in the interoceptive field; but they are stimulated by outside agents, therefore they are considered as forming part of the exteroceptive field, because they give information about the external environment. Smell has no proper localization; nevertheless by integration with other sensations it is "projected" outside the body. Taste has contact receptors.

Contact receptors. Tactile, temperature, and pain receptors give rise to sensations that are localized on the stimulated receptor; often they have a high affective tone. The response is usually of a terminal nature and completes adaptation.

Proprioceptors. Receptors placed in the depth of tissues, muscles, tendons, etc., evoke less well-defined sensations. The reactions are usually unconscious and are of limited scope, such as the myotatic reflex. Labyrinthine reflexes, which arise in proprioceptive receptors, involve, however, a large number of muscles widely distributed in the organism.

Visceroceptors. Receptors on the internal surface of the body give rise to poorly defined and badly localized sensations. The most important reactions initiated in these receptors take place in the viscera (viscerovisceral reactions) and are usually unconscious. Other responses are carried out by somatic effectors (viscerosomatic reactions), and sometimes the sensitiveness of the outer surface is modified (viscerosensitive reactions).

STIMULATION OF RECEPTORS

Müller's law of specificity of nervous energy. Excitation of a receptor always gives rise to the same sensation, whatever the nature of the stimulus. Thus stimulation of the retina by light gives a sensation of light; stimulation by a mechanical or electrical stimulus also evokes a sensation of light. This fact, known as "Müller's law of specificity of nervous energy,"

[1] SHERRINGTON, *op. cit.*, p. 325.

was clearly stated by the famous German physiologist a century ago. This is due to the termination of the afferent path in the cerebral cortex, as is demonstrated by stimulating any part of the afferent path or the cortical center. Thus stimulation of the optic nerve or pathway, or of the occipital cortex, gives rise to a luminous sensation. This also explains why, after amputation of a limb, the patient refers irritation of the severed nerve endings, and pain thus caused, to the nonexistent limb.

Specific irritability of the receptors. There is an adequate stimulus for each receptor, *i.e.*, a form of energy for which it has a low threshold, while the threshold is high for other stimuli; this is known as the specific irritability of the receptors. The appropriate stimulus for the photoreceptor is light, *i.e.*, electromagnetic waves of 330 to 900 mμ wavelength. Vision is limited on the red side of the spectrum because rhodopsin (visual purple) does not absorb rays of more than 800 to 900 mμ. Ultraviolet rays do not stimulate the retina; they produce fluorescence in the transparent media of the eye. Mechanical vibrations of frequencies between 16,000 and 20,000 per second stimulate the human phonoreceptors. Certain substances in solution are the specific stimuli of the so-called "chemical receptors," *i.e.*, smell and taste. Deformation of the skin stretches or deforms the tactile receptors and acts as their specific stimulus. There are two kinds of temperature receptors; some respond to cooling, and others to warming, of the skin. Any harmful agent that damages the cells stimulates the pain receptors, probably because the damaged cells set free substances (algogenous substances) that irritate the free nerve endings in the tissues. Proprioceptors respond to pressure, stretching, and changes in tension.

QUANTITATIVE RELATIONS BETWEEN STIMULUS AND RESPONSE

Strength of response, measured objectively, *e.g.*, by electrical registration, or subjectively by intensity of sensation, is related to several parameters of the stimulus. Thus there are thresholds for intensity, time, area on which the stimulus acts, and several difference thresholds.

Intensity threshold. The amount of energy required to stimulate a receptor depends on several factors: (*a*) the nature of the stimulus; (*b*) the condition of the receptor; (*c*) the locus of stimulation; (*d*) the species.

The nature of the stimulus. The threshold is low for the specific stimulus and high for other stimuli, *i.e.*, a small quantity of energy is efficient in the first case, while large quantities are needed in the second.

The visual threshold for light has been measured in optimal conditions of adaptation to darkness, with very brief stimuli, of selected wavelength, applied on a small surface of the peripheral retina. For a blue-green light it is 2.1 to 5.7×10^{-10} erg. The 500 rods within the stimulated area absorb 5 to 14 quanta of energy; therefore either a single rod absorbs the total amount, or 5 to 14 rods absorb 1 quantum each. The visual threshold is, therefore, close to the limit imposed by the nature of light, because the receptive cell apparently responds to the smallest particle of energy that exists.[1]

In complex receptors, such as the eye, there are differences in the thresholds for the various kinds of specific stimuli. Thus in the light-adapted eye the green-yellow part of the spectrum is the most luminous. The threshold sensation is produced by 0.0015 watt per lumen (see Chap. 76) of light of 550-mμ wavelength; with other wavelengths greater quantities, in some cases many thousand times this amount, are needed.

The condition of the receptor. The intensity threshold of a receptor does not remain at a constant level but varies in different circumstances. If stimuli very near the threshold strength are applied, the intensity of the sensation varies from one moment to another, as if spontaneous fluctuations in sensitiveness occurred.[2] There are many factors that cause variations in threshold; the following can be mentioned:

1. Simultaneous or previous application of a stimulus raises the threshold for certain stimuli and lowers it for others (see further

[1] Hecht, S., S. Shlaer, and M. H. Pirenne, *J. Gen. Physiol.*, **25**, 819, 1942.

[2] Apparently spontaneous fluctuations in the threshold of visual sensations may be due to the stimulus, because the threshold of this receptor is so low that it is statistically probable that the receptor does not always absorb the same quantity of energy. On the other hand, fluctuations are also observed in auditory sensations. In this case the amount of energy that must be used to obtain a response is such that the influence of the stimulus can be discarded; therefore fluctuations must be due to variations in the sensitiveness of the receptor.

on, "Simultaneous and successive contrast" and "Masking").
2. Adaptation takes place, *i.e.*, certain specific changes in the receptor can lower or raise the threshold considerably; *e.g.*, a dark-adapted eye is sensitive to approximately 0.0000007 millilambert, using achromatic light, while the eye adapted to a luminosity of 2,000 millilamberts is not sensitive to light intensities below 4 millilamberts.
3. Metabolic conditions are also of importance; *e.g.*, deficiency in vitamin A interferes with the formation of visual purple and diminishes sensitiveness to achromatic light. Anoxia and hypoglycemia also increase the threshold of a dark-adapted eye.[1]
4. In old age, sensory acuity diminishes.

The locus of stimulation. A tactile sensation can be obtained with a weight of 2 gm. per sq. mm. on the tip of the tongue; on the palm of the finger tip, 38 gm. per sq. mm. is necessary, and 48 gm. per sq. mm. on the skin of the loins. The different parts of the retina also have different thresholds. This is due to uneven distribution of receptor cells or to variations in the type of these cells in different parts of the receptive field.

Species differences. Sensory capacity varies in different species. Two methods have been used to measure objectively and accurately the sensitiveness of receptors in animals: (*a*) establishment of conditioned reflexes; (*b*) registration of electrical activity evoked by excitation of the receptor. Thus it has been proved that rats are less sensitive than man to sounds of frequencies below 8 kc. (kilocycles) but are more sensitive to sounds of a higher pitch. The lowest threshold is for 20 kc., a pitch almost inaudible to the human ear, and sounds of 40 kc. are well perceived. Bats have an even greater sensitiveness; they can hear ultrasounds of over 100 kc. Vision in man is more sensitive than in the dog, an animal that is almost color-blind. Visual acuity is similar in man, the chimpanzee, and the rhesus monkey; this monkey can distinguish colors almost as well as man, except in the red part of the spectrum, to which it is less sensitive. Many other data on the sensory capacity of animals have been obtained; some of them will be mentioned further on.

[1] Subjects breathing 13 per cent oxygen and injected with 4 units of insulin showed a fourfold increase in threshold (McFarland, R. A., and W. H. Forbes, *J. Gen. Physiol.*, **24**, 69, 1940).

Time threshold. Two time factors must be considered in connection with the stimulation of receptors: (*a*) the time in which the energy change takes place; (*b*) the duration of the constant phase of the stimulus. If the gradient of change is not sufficiently steep, the receptor adapts itself before the threshold can be reached. Utilization time is related to the intensity of the stimulus in the dark-adapted eye ($T \times I = K$) for times less than 0.1 sec. Chronaxie determinations confirm this. In complex receptors there is more than one chronaxie, as each type of cell has its own; thus the peripheral retina has a chronaxie of 1 to 1.9 msec., corresponding to the rods, and the fovea a chronaxie of 2.1 to 2.8 msec., corresponding to the cones.

Area (or size) threshold. A stimulus that acts on a circumscribed area of a receptive field stimulates only one sensory unit or a small number of them; on the other hand, a widespread stimulus excites a large number of units, and the impulses they evoke are summated in the center. Threshold intensity, therefore, decreases as the size of the stimulated surface increases. For example, a source of light must be of a certain size, which must increase as luminosity decreases, for the eye to perceive it. For achromatic light,

$$S \times I = K$$

where S is the area of stimulated surface and I is the luminosity. A tactile sensation can be produced on the skin by light touches if these act on a large surface, but the stimulus must be of greater intensity if it acts on a more limited area.

Difference thresholds. It is possible to discriminate the strength, duration, and other properties of different stimuli. The minimum difference that can be perceived is known as the difference threshold of intensity, time, etc.

Difference threshold of intensity. Weber examined the capacity of the tactile receptor and concluded that to obtain an increase or decrease in sensation it was necessary to increase or decrease the stimulus by a definite fraction of its total strength (Weber's law). The ratio of change varied between 1:10 and 1:40 in different parts of the skin. Thus if a stimulus of 0.9 gm. is distinguished from one of 1 gm. (a ratio of 10 per cent), a stimulus of 10 gm. must be increased or decreased by 1 gm. to produce a stronger or

weaker sensation. The eye distinguishes differences in luminosity of 1:167 (Helmholtz). The ear can differentiate from 1:9 to 1:20 in sound intensity. Fechner has stated Weber's law with greater precision (Fechner's law): above the threshold and up to a certain maximal intensity, sensation is a linear function of the logarithm of the stimulus (Fig. 385). The condition of the receptor modifies its discriminating capacity. Thus when the skin temperature is between 27 and 32°C., differences of 0.2°C. can be distinguished, but at lower or higher temperatures discrimination is not so efficient.

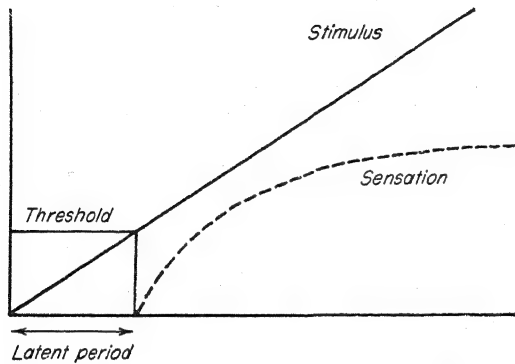



FIG. 385. Fechner's law. Solid line, size of stimulus; broken line, intensity of sensation.

Difference threshold of time. Successive stimuli can be perceived separately if the interval between them is not too short. When the interval is below a minimum, the sensation becomes continuous. The minimum frequency necessary to produce fusion varies from one receptor to another. It is influenced by the condition of the receptor and by the nature and strength of the stimulus. Adaptation (see further on) plays an important part in fusion. Successive mechanical stimuli can be individually perceived up to a frequency of 500 to 600 per second, but faradic stimulation becomes continuous at rates above 120 to 150. The eye has a slow adaptation rate, and threshold stimuli must be repeated at least every 0.25 sec. to avoid flicker; as the intensity of the stimulus increases, the interval must be reduced to obtain complete fusion.

Difference threshold of size. Differences in the size of the surface stimulated can also be perceived. Thus the eye distinguishes the size of objects when they differ by 1:100 if they are compared simultaneously in optimal conditions. Discrimination is less efficient if the objects are seen one after the other. Simultaneous stimulation of two points on the skin gives rise to two separate sensations if the distance between the points exceeds a minimum, which varies from one part of the body to the other (1 mm. on the tongue, 67 mm. on the back). Discrimination occurs only when stimuli act on separate sensory units. On the retina (central vision) the images must fall on separate cones (visual acuity); on the skin, receptors innervated by at least two afferent neurons must be stimulated.

In addition to the difference thresholds already mentioned, there are others, which also condition the analytical capacity of the senses. Thus the ear not only differentiates pitch and quality (timbre) in a sound, but it can also pick out the melody played by a certain instrument in an orchestra and distinguish it from the rest of the symphony. Vision distinguishes brightness, color (wavelength), and color saturation (the amount of white light that "dilutes" the color).

There are considerable individual differences in discriminating capacity. For example, the normal eye can differentiate about 165 chromatic tones; in the yellow region differences of 1 $m\mu$ are perceived, whereas in the red and purple, the variation must be of 6 to 8 $m\mu$ to be appreciated. Certain individuals, however, distinguish fewer tonalities and others are more or less color-blind (see "Color vision," Chap. 76). Experience and education considerably increase the capacity to discriminate; a painter or a musician will perceive fine differences that are not noticed by those who lack an artistic or a musical education.

THE RECEPTOR OR SENSORY UNIT

The motor unit consists of a motor neuron and all the muscle fibers innervated by the branches of its axon; likewise, a sensory unit is made up of an afferent neuron and all the receptors innervated by its peripheral branches.[1] The distribution of these units in a receptive field has been studied by anatomical methods and by registering the electrical potentials in single afferent fibers.[2] For example, in the central part of the fovea in the retina, each bipolar cell is connected peripherally with only one cone and centrally with one ganglion cell. In the peripheral parts of the retina each bipolar

[1] TOWER, S. S., *J. Neurophysiol.*, 3, 486, 1940.

[2] ADRIAN, E. D., and J. ZOTTERMAN, *J. Physiol.*, 61, 151, 1926; ADRIAN, E. D., McK. CATTELL, and H. HOAGLAND, *J. Physiol.*, 72, 377, 1931.

cell innervates several cones or rods; cone units are less extensive than rod units. In the rabbit's skin one afferent fiber innervates up to 300 hair follicles, but each follicle is innervated by at least two nerve fibers and sometimes by as many as seven.[1] The tactile units in the cat's tongue have a diameter of 5 mm., and those of the frog's skin cover 4 to 10 sq. mm. at least. The area of a pain unit has a diameter of 0.75 mm. on the back of the hand in man; it is much smaller on the palmar surface of the fingers, and twice as large on the skin of the forearm. Sensory units on the surface of the frog's viscera cover an area of 2 to 3 sq. cm. or more. On the cornea a unit covers a quadrant and spreads out into the neighboring sclera and conjunctiva, but it does not innervate the whole region in a uniform manner. These examples show that the size of the unit varies in the different receptive fields and sometimes in different parts of the same receptive field.

The receptive fields of neighboring sensory units overlap, so that a given point may be innervated by two or more units. This is demonstrated by cutting a cutaneous nerve. If the nerve is sufficiently large, there will be a region of complete anesthesia, surrounded by another of hypoesthesia, in which sensitiveness increases from the anesthetized to the normal parts. The intermediate region is due to suppression of units belonging to the cut nerve and persistence of units belonging to neighboring nerves that have remained intact.

[1] WEDDELL, G., *J. Anat.*, **75**, 346 and 441, 1941.

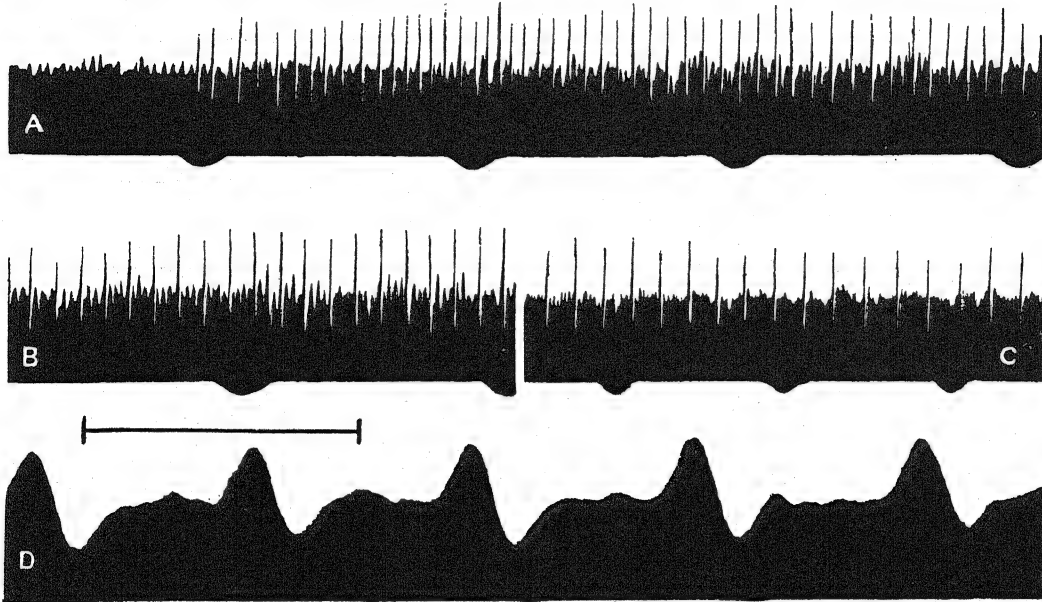



FIG. 386. Impulses in a single afferent fiber from a proprioceptive receptor in the toe of a frog. *A*, stimulation by stretching the toe with a load of 5 gm.; *B*, 2 sec. after the beginning of stimulation; *C*, 5 sec. after the beginning of stimulation; *D*, at high speed, the outline of the electric variation is invariable. (*Courtesy of Professor E. D. Adrian.*)

Stimulation of single receptors. Adrian and his collaborators studied the action potentials of an afferent nerve after stimulating a single receptor in the field it covered. Responses from a single receptor were also obtained by cutting all but one or two fibers of a nerve and then stimulating the area innervated by this nerve. The first experiments were made on a single proprioceptor of the sternocutaneous muscle of the frog[1] and of a muscle of the middle toe of the same animal;[2] both were stimulated by stretching. Later, single receptors were stimulated in many other receptive fields

A series of impulses follows the application of a single stimulus to a receptor, not merely a single response such as occurs when the nerve is stimulated (Fig. 386). The speed with which the frequency of the discharges reaches a maximum increases with the strength of the stimulus. The maximum frequency also increases in relation to the strength of the stimulus,

[1] ADRIAN, E. D., and J. ZOTTERMAN, *J. Physiol.*, **61**, 161, 1926.

[2] MATTHEWS, B. H. C., *J. Physiol.*, **67**, 169, 1929.

following an exponential curve. For example, if a proprioceptor is stimulated by stretching a muscle, the maximum frequency of discharge increases as the logarithm of the weight used as stimulus. This is in agreement with Fechner's law. The maximum frequency is limited by the refractory period, which is slightly greater than that of the afferent nerve. Factors that modify the refractory period, such as fatigue which lengthens it and heat which shortens it (in the frog), also modify the maximum possible frequency.

Continuous stimulation evokes a discontinuous response. After the first impulse has been fired off a second impulse is discharged as soon as the receptor recovers its excitability sufficiently to respond to the stimulus that is still acting. The threshold is lower for a strong stimulus than for a weak one; therefore it is reached sooner, and frequency of impulses increases with the strength of the stimulus. When impulses are discharged at low rates, the intervals between them are usually irregular; in this case the refractory period has ended, and there is complete recovery between two impulses. The rate of discharge in physiologic conditions is far below the maximum possible.

ADAPTATION

The rate at which a receptor discharges diminishes even when it is stimulated continuously. This decrease is particularly marked in touch receptors, which rapidly cease to discharge; it is less marked in pacinian corpuscles. This is known as adaptation. It occurs at a faster rate with strong stimuli than with weak ones; also after a receptor has been repeatedly stimulated. Adaptation is almost lacking in stretch receptors, which continue to discharge at a low rate for a considerable time (Fig. 387).

Several factors modify adaptation; *e.g.*, the ion equilibrium. Thus, a decrease in Ca^{++} (by means of oxalate) delays adaptation in the touch receptors of frogs. If the calcium concentration falls even more, spontaneous bursts of activity are observed. An increase in K^{+} has the same, while an increase in Ca^{++} produces the opposite, effects.

Adaptation occurs in the receptor itself and not in its nerve fiber. Thus adaptation of a receptor in a sensory unit has no influence on that of other receptors innervated by branches of the same axon. Adaptation may be due in some cases to a progressive loss in the efficiency of the stimulus. For example, mechanical deformation stimulates the tactile receptors on the hair follicles and the pacinian corpuscles; in the hair follicles adjustment to the new mechanical conditions takes place rapidly and the receptor soon ceases to discharge, while it is more gradual in the pressure corpuscles, which therefore continue to discharge for a longer time.

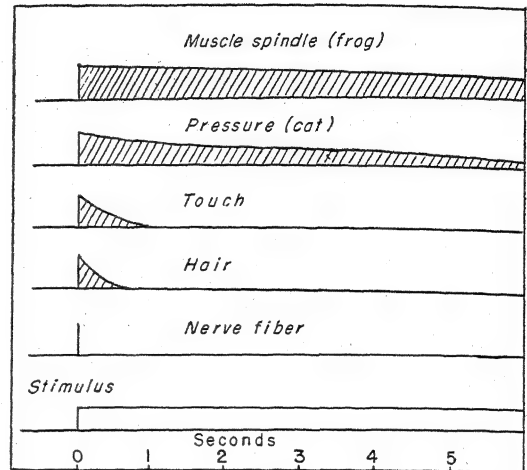



FIG. 387. Adaptation in different receptors. Abscissa, time in seconds. The stimulus is applied continuously. Frequency of discharge from a muscle spindle (top curve) remains high throughout the period of observation; that of a pressure corpuscle (second curve) falls gradually. Tactile corpuscles (third curve) and touch receptors of hair roots (fourth curve) cease to discharge in more or less 1 sec. The nerve fiber (last curve) responds with a single impulse. (*After Adrian.*)

After-discharge. Certain receptors continue to discharge impulses after stimulation has ceased. This is known as after-discharge. It is more marked in proprioceptors, less in pain receptors, and almost nonexistent in touch receptors.

Fatigue. Adaptation must not be mistaken for fatigue. When the receptor is fatigued by repeated and prolonged stimulation, the initial frequency of discharge diminishes and adaptation is more rapidly established. Fatigue occurs sooner with strong and prolonged stimulation, and disappears after a period of rest. Anoxia accelerates the onset of fatigue and retards recovery, but it has no influence on adaptation.

Intensity of sensation. When the strength of the stimulus is increased, the receptors discharge at a higher frequency and more sensory units

are activated. Thus intensity of sensation is conditioned by two mechanisms: (*a*) the rate with which the impulses arrive to the center along a given afferent path; (*b*) the number of afferent paths in activity.

Latent period and reaction time.[1] The latent period of sensation, *i.e.*, the interval between the application of a stimulus and its perception (perception time) can be measured only indirectly. Stimulation of receptors disturbs the α-rhythm in the electroencephalogram of a resting subject (see Chap. 87). This fact has been used to measure the time taken by a stimulus, *e.g.*, a painful stimulus, to produce an effect on the cortex.[2] Since perception is dependent on cortical activity, this method gives a near approach to perception time.

Usually the reaction time is measured, *i.e.*, the interval between the application of a stimulus to a receptor and the response of the subject on perceiving the stimulus. The reaction time to a given stimulus varies considerably from subject to subject, and in the same subject from day to day. The site of the receptive surface on which the stimulus is applied also has an influence on the reaction time. Reaction time to stimulation of a cutaneous tactile receptor diminishes with increasing strength of stimulus and with increasing area of surface stimulated. The relationship in both cases is a logarithmic one,[3] *i.e.*, it follows Fechner's law. Fatigue increases reaction time; repetition of the stimulus, *i.e.*, training, diminishes it.

An important fact made evident by these studies is that *whatever recebtor is stimulated, and whatever stimulus is used, the effect on the afferent nerve is fundamentally the same:* impulses are discharged which can vary in number, rate, voltage of the action potential, speed of conduction, and other details, but they are essentially alike.

SIMULTANEOUS AND SUCCESSIVE CONTRAST

Stimulation of a receptor diminishes the sensitiveness of neighboring parts of the receptive field for the same kind of stimulus and augments it for the opposite kind of stimulus. This explains simultaneous contrasts, which are easily demonstrated in color vision. If a red disk is placed on a white surface, a green halo is seen around it, because the threshold for green is lowered and that for red increased in the proximity of the parts of the retina stimulated by red light, and the green component of white light can be perceived.

After a stimulus has ceased to act on a receptor, the threshold is increased for this kind of stimulus and lowered for the opposite kind. Thus a stimulus when it follows one of an opposite kind evokes a more intense sensation than when it is applied alone (successive contrast). In the example already given, on removing the red disk from the visual field a green patch of the size and shape of the disk is seen on the white surface.

Contrast phenomena are also observed in other sensations. For example, if one hand is submerged in water at 10°C. and the other in water at 38°C., and then after a few minutes both hands are placed in a basin with water at 25°C., the water will seem warm to the hand that has been at 10°C. and cold to the one that has been at 38°C. The receptors for warmth have become hypersensitive in one hand and those for cold in the other.

LOCALIZATION

Sensations have a local sign; they are referred either to some point on the receptive surface, which is not always the place stimulated, or to some place outside the organism, related to the source of the stimulus. In the skin there are spots that are specifically sensitive to different kinds of stimuli (Blix, von Frey), and in the cerebral cortex the different parts of a receptive field have a localized representation. These two facts led to the belief that the local sign of a sensation is due to transmission of the impulses from a receptor, through a definite chain of neurons, to a circumscribed and fixed area in the cerebral cortex. This correlation of points in the periphery with points in the cerebral cortex, through a chain of neurons, exists in the central part of the fovea of the retina, for which there is point-to-point projection in the lateral geniculate body and the occipital cortex. There is, however, a certain degree of overlapping (*i.e.*, the terminals of a given axon end on the somas of more than one neuron, and each soma receives terminals from more than one axon) even in the pathway of foveal vision,[1] which has a very high degree of spatial discrimination.

[1] MICHON, P., "Le Temps de réaction," Masson et Cie, Paris, 1939.

[2] GORDON, G., and D. WHITTERIDGE, *Lancet*, **1**, 700, 1943.

[3] LELE, P., D. C. SINCLAIR, and G. WEDDELL, *J. Physiol.*, **123**, 187, 1954.

[1] O'LEARY, J. L., *J. Comp. Neurol.*, **73**, 405, 1940.

Cutaneous and visceral sensory unit fields have been found to overlap and the same probably occurs in other receptive fields. There is convergence and divergence in the centers at all levels. Afferent nerve impulses are, therefore, not transmitted from the receptor to the cortex along a simple chain, each link of which is a neuron, but through complex ramified multineuron chains, in which are interposed closed reverberating circuits of excitation (Fig. 388).

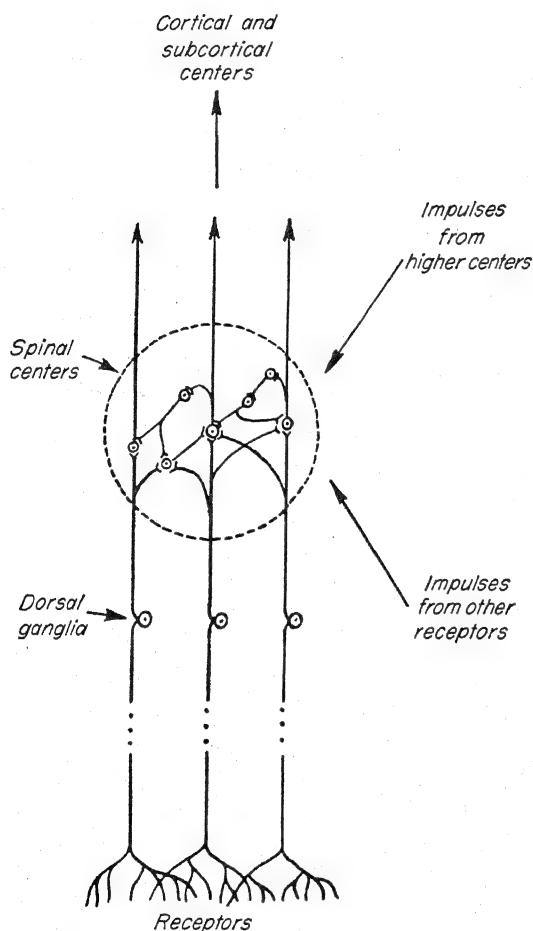



Fig. 388. Diagram of sensory nerve path (highly schematic). There is overlapping of sensory units in the periphery, and convergence, divergence, and internuncial neuron circuits in the spinal cord. A similar architectural plan is found in cortical and subcortical centers.

Functional architecture undoubtedly conditions the local sign of reflexes and sensation, but not in such a simple way as was at one time supposed. Impulses arrive at the centers and eventually at the cortex, not from a single receptor but from several receptors stimulated simultaneously or successively more or less intensely. An excitation pattern is formed which is analyzed at the center, recognized, and localized as a "sensory point."[1] A weak stimulus may activate only one or a few receptors of a single sensory unit; a strong stimulus excites receptors of several units with different intensity. In the first case the sensation is referred to the whole field covered by the unit; in the second, localization is more precise because the overlapping area diminishes in size as the number of units that serve it increases (Fig. 389). Accuracy of localization depends therefore on (*a*) the size of the sensory unit field; (*b*) the number of overlapping units covering a given field; (*c*) the number of units activated, *i.e.*, the strength of the stimulus. Experience, *i.e.*, repeated application of a stimulus, is of considerable importance for accuracy of localization.

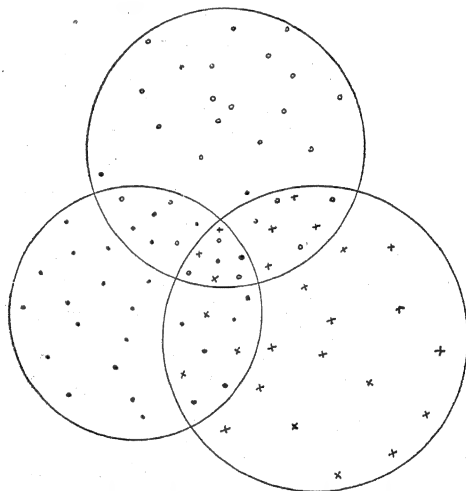



Fig. 389. Localization. Diagrammatic representation of sensory fields of three overlapping units. When the receptors of only one unit are stimulated, the sensation is referred to the whole area covered by the unit. When two or three units are stimulated, the sensation is referred to the overlapping areas.

BIBLIOGRAPHY

See page 902.

[1] Tower, S., *Assoc Res. Nerv. Ment. Dis.*, **23**, 16, 1943.

CHAPTER 72

Cutaneous Sensation. Touch, Temperature Senses, and Pain

THE SKIN IS a receptive surface in which there are many nerve endings. In man, these nerve endings are of two types (Fig. 390): (*a*) fine, profusely ramified, nonmyelinated free terminals arising in myelinated or in large nonmyelinated fibers; and (*b*) complex, closely knit, localized endings, surrounded or not by a visible capsule of variable thickness. Among the latter, several different types have been described: the Meissner corpuscles, Krause's end-bulbs, Ruffini's endings, and Merkel's disks, to which can be added the nerve baskets of the hair follicles. Many years ago Ruffini[1] remarked on the great variability of peripheral nerve endings and the difficulty in classifying them. More recent work has confirmed this opinion, showing that there are many intermediate forms "from the simplest to the most complicated, from superficial loosely encapsulated whorls, to the thickly capsulated

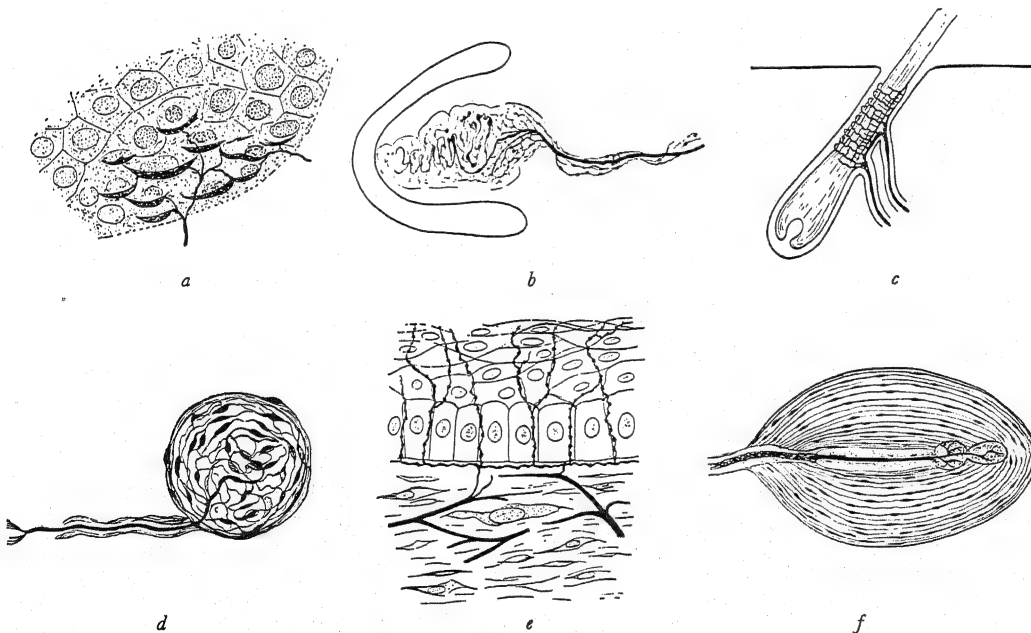



FIG. 390. Cutaneous receptors. *a*, Merkel's disks; *b*, Meissner's corpuscle; *c*, nerve basket of hair follicle; *d*, Krause's end-bulb; *e*, free nerve endings in the cornea (*Cajal*); *f*, pacinian corpuscle.

[1] RUFFINI, A., *Rev. gén. d'histol.*, **1**, 421, 1905.

highly complex dermal corpuscles."[1] These "organized nerve endings" are found in the mucous membranes of the mouth, genitals, anus, and the conjunctiva, and in hairless parts of the skin, such as the plantar surface of the foot and the palmar surface of the hand and fingers. They are distributed in two groups, one immediately beneath the epithelium and the other in the dermis 0.5 to 1.5 mm. below the surface. The more deeply placed corpuscles usually have a thicker, apparently noncellular, capsule. In hairy parts of the skin these organized nerve endings have not been found; there are only nerve baskets around the hair follicles and free nerve endings.

The end organs are usually distributed in groups. Each one receives one or two myelinated nerve fibers, which lose the myelin sheath on penetrating into the corpuscle; one nonmyelinated fiber also innervates a group.[2] The nerve baskets of the hair follicles are formed by two to seven nerve fibers, and the same axon sends branches to several follicles. This type of receptor is highly developed in the vibrissae of nocturnal animals, such as cats and rodents.

There are four main modalities of sensation, *i.e.*, touch, cold, warmth, and pain. Several other distinct types of sensation are recognizable, but these are usually considered as produced by special patterns of activation of the nerve endings corresponding to the main modalities, or the combined stimulation of receptors of two or more modalities at various frequencies and intensities. Thus itch is a special type of response to weak stimulation of pain receptors, and vibratory sensation is produced by intermittent stimulation of touch and pressure receptors. Other sensations such as tickling, roughness, smoothness, wetness, tridimensional form (stereognosis), etc., are also complex sensations. "The selection of four fundamental modalities for the classification of cutaneous sensation is a compromise between simplicity and confusion, and between specificity of stimulation and flexibility of central interpretation."[3]

The work of von Frey, Goldscheider, and Blix demonstrated that the skin is a mosaic of sensitive "spots" (punctate sensibility), *i.e.*, that certain parts are highly sensitive to one kind of stimulus, while the areas surrounding them are relatively insensitive. Each modality of sensation has been attributed to a special type of receptor. There is good evidence that this is so in the case of deep sensibility; the muscle spindles and Golgi tendon organs respond to differences in tension, and the pacinian corpuscles mediate pressure sensations. In the skin, movement of a hair stimulates the nerve basket surrounding its follicle, and shaving considerably diminishes sensitiveness to touch. Meissner's corpuscles and Merkel's disks have been considered touch receptors, Krause's end-bulbs cold receptors, and Ruffini's corpuscles warmth receptors. The evidence, however, is far from satisfactory. As has already been mentioned, identification of the different types of end organs is not easy, and some of them may be only artefacts, *e.g.*, Merkel's disks, because these structures are very sensitive to manipulation. Histological examination of the skin where "spots" for touch, cold, or warm sensation had been localized by previous stimulation has not given consistent results. Moreover, cold and warm sensations can be evoked from parts of the skin (lips, ear, abdomen) where only free nerve endings and nerve baskets of hair follicles have been found. Therefore, no exclusive association between a specific form of nerve ending and a specific modality of cutaneous sensation can so far be considered as certain.

The sensibility of the skin varies considerably from one part to another, depending on the number of receptors. A high degree of sensitiveness to one modality of sensation does not necessarily coincide with an equivalent degree of sensitiveness to another. Stimulation of a single spot, however, may arouse different sensations according to the type of stimulus applied, either because the spot contains receptors corresponding to two or more modalities, or because different patterns of stimulation are applied to the same receptor; *e.g.*, weak stimulation of pain endings provokes "nonpainful" sensations, such as itch.

TOUCH

Touch is the modality of sensation aroused by mechanical deformation of the skin. Meissner showed that, when a finger is submerged in mercury, a sensation is felt only at the level of

[1] Sinclair, D. C., G. Weddell, and E. Zander, *J. Anat.*, **86**, 402, 1952; Hagen, E., H. Knoche, D. C. Sinclair, and G. Weddell, *Proc. Roy. Soc., London, s.B.*, **141**, 279, 1953.

[2] Weddell, G., *J. Anat.*, **75**, 441, 1941.

[3] Bishop, G. H., *Physiol. Rev.*, **26**, 77, 1946.

the surface between air and mercury, where there is deformation of the tissues, and not over the rest of the finger, where there is uniform pressure. Touch cannot always be clearly distinguished from light pressure, *i.e.*, the persisting sensation aroused by the application of mechanical force to the skin over a wider area than that necessary to provoke touch sensations. There is therefore considerable ambiguity in the interpretation of touch and light pressure sensations. Some authors consider that they are different stimulation patterns of the same modality of sensation (touch-pressure); others, that they are two separate modalities. Pressure in this case is not considered a cutaneous sensation in the strict sense of the term, because it is mediated by the pacinian corpuscles situated in the inner layers of the dermis, the subcutaneous tissues, and other deep structures, such as the subperitoneal tissue, tendons, periosteum, etc.

FIG. 391. Cutaneous unit areas. Above, *A* and *C*, unit touch areas with strong stimulus; *A′* and *C′*, with weak stimulus; *C″*, with threshold stimulus. *B*, unit touch areas with 0.0001 mf. stimulus; hair shafts; high spot for touch. Three of the areas overlap; above the fourth area there is a space insensitive to touch with three hair shafts sensitive to touch. E_1, single touch area containing no hair shafts; E_2, three touch areas with · high spot in two of them to which the sensation is referred; E_3, area around isolated hair shaft (*hs*); *pp*, pain points; *tp*, touch point; $\odot$, hair endings. Stimulation of the touch point with the threshold for touch also stimulated the adjacent pain point. Over the pain point itself the pain threshold was much lower. (*Bishop, G. H., J. Neurophysiol., vol. 6, p. 361, 1943.*)

Sensory thresholds. There are several methods for stimulating touch receptors. One of the oldest is von Frey's, which consists in touching the skin with a hair mounted on a wooden handle. The length and thickness of the hairs vary so that different weights are needed to bend them, and the force exerted on the skin can thus be measured. Hensen used glass-wool thread instead of hair; nylon threads of standard thickness have also been used. When touch is explored by lightly stroking the skin with the loose strands of a wad of cotton wool (Head's method) or a camel-hair brush, only qualitative results are obtained. Bishop[1] has used high-voltage low-current sparks to stimulate single sensory units without deforming the skin. By means of this method he has mapped out the areas covered by touch and pain units in regions with many receptors of different types. This method was combined with anesthesia of small cutaneous nerve branches. Small units were thus isolated in which a single sensitive point was found surrounded by an insensitive area[2] (Fig. 391). The size of the area covered by one unit varies in different species and in different parts of the skin, *e.g.*, 5 sq. mm. in the cat's tongue and 100 sq. mm. in the skin of the frog. There is considerable overlap among the areas of different units.

[1] BISHOP, G. H., *F. Neurophysiol.*, **6**, 361, 1943.
[2] BISHOP, G. H., *F. Neurophysiol.*, **7**, 71, 1944.

Certain regions have a very low touch threshold, *e.g.*, the tip of the tongue, lips, nose, and the balls of the fingers. The dorsal aspect of the hand and forearm have a considerably higher threshold than the palm and the anterior aspect of the forearm. The skin of the back, loins, and lateral aspect of the thighs has a very high touch threshold. The number of receptors is much larger and the size of the sensory unit much smaller in the more sensitive parts than in the less sensitive ones.

Difference thresholds. Regions that have a low intensity threshold also have low difference thresholds. A difference in intensity of sensation can be perceived on the lips and forehead if the strength of the stimulus varies by one-fortieth or one-thirtieth, while the change must be of one-tenth to be recognized on the thigh, shin, or foot.

Spatial discrimination (two-point discrimination) is explored by means of Weber's compass. When its two blunt points are applied to the skin, they are separated by a distance sufficiently large for them to be perceived as separate points of contacts. The distance is then reduced until a single sensation of contact is aroused. Spatial discrimination varies considerably from one region to another. Areas with a low intensity threshold also have a low two-point threshold, but variations in the two thresholds are not strictly parallel (Table 93). Spatial discrimination can be increased (*i.e.*, the distance between two points at which a double sensation is perceived can be reduced) by

Table 93. Distance between Two Stimuli Needed to Obtain Separate Touch Sensations

Area	*Distance, mm.*
Tip of tongue	1.1
Palmar surface of the finger	2.3
Back of the third phalanx	6.8
Palm	11.3
Back of the hand	31.6
Neck	54.0
Arm, thigh, back	67.1

searching out touch spots. Thus on the ball of the fingers it is only 0.1 mm. in some places. The two-point sensibility of the skin is, however, far below that of the eye.

Fusion, i.e., a continuous sensation, is not aroused by intermittent mechanical stimulation at rates below 500 per second; each stimulus is perceived separately. On the other hand fusion occurs at rates between 130 and 150 per second when induction shocks are applied to the skin.

Fatigue and old age diminish sensitiveness to touch, especially the capacity to discriminate.

Adaptation. Touch receptors show a high degree of adaptation and have no after-discharge. Thus, nerve baskets around the hair follicles cease to discharge after a few seconds of continuous stimulation. This fact explains in part why the clothes are not felt unless they are displaced by movement and thus excite receptors that have been quiescent. Nerve baskets of the vibrissae of nocturnal animals do not adapt so rapidly, and they continue to discharge impulses during several minutes.

Localization. Touch sensations are fairly accurately referred to the point stimulated (within 1 cm. on the forearm). Localization is conditioned by experience. Aristotle's observation is a remarkable demonstration of this fact; if two fingers are crossed so that the sides normally separated face each other, an object that touches both of them at the same time is felt as if it were double. Accuracy of localization and discrimination can be increased by training, as occurs in blind people. There is also, to a certain extent, the capacity to "project" the sensation outside the organization, as when a surgeon explores a cavity with a probe. The affective tone of touch is very high and usually of a pleasurable nature. Rapid adaptation prevents intense stimulation from becoming disagreeable.

TEMPERATURE SENSES

There are two temperature senses; certain spots in the skin are sensitive to cold, others to warmth. Thermal receptors are sensitive to the temperature of the skin and its changes, provoked by variations in the temperature of the environment or changes in the circulation of the skin; thus, blushing arouses a sensation of warmth and cutaneous vasoconstriction one of cold.

Bazett and his associates[1] have measured with thermocouples the rate of penetration of heat through the skin and correlated it with latency of sensation. They conclude that receptors for

[1] Bazett, H. C., Temperature Sense in Man, in "Temperature, Its Measurement and Control in Science and Industry," Reinhold, New York, 1941.

cold are placed superficially and those for warmth more deeply, approximately at the levels at which Krause's and Ruffini's corpuscles, respectively, are found. In the cat's tongue cold receptors are placed close beneath the epithelium, as has been shown by measuring the latency of spike potentials evoked in cold fibers in the lingual nerve and simultaneously recording the temperature changes in the tongue.[1]

There are spots for cold and warmth in the skin. The former are more numerous than the latter; it has been estimated that there are 250,000 cold spots and 30,000 warm spots. Cold spots are distributed in clusters, covering areas with irregular contours, which are separated by zones where there is little or no sensitiveness to cold. In man there are more cold spots per unit of skin surface in the thorax and upper limbs than on the lower limbs.[2]

Many factors, such as changes in the circulation and in the temperature of the skin, alter the apparent number of receptors. Thus, on raising the temperature of the skin from 25 to 33°C. the cold spots were found to increase from four to five times.[3] This has been attributed to an increase in the number of receptors that respond at the higher temperature, or to a higher rate of discharge from the receptors, which would thus reach the central threshold of sensation. Local injections of acetyl-β-methylcholine or prostigmine increase, and of adrenaline or atropine decrease, the number of cold spots significantly.[4] Acetylcholine injected into the cat's tongue increases the rate of constant discharge of cold receptors and their response to a fall in temperature.

Stimulation of receptors. Thermal receptors can be stimulated by means of tubes of small bore through which water at different temperatures is kept circulating. More precise methods of stimulating with constant temperatures or changes in temperatures, and recording the temperature of the stimulated area with thermodes, have been devised.[5] Receptors for

[1] Hensel, H., L. Ström, and Y. Zotterman, *J. Neurophysiol.*, **14**, 423, 1951.
[2] Carlson, L. D., *Proc. Soc. Exper. Biol. Med.*, **85**, 303, 1954.
[3] Bing, H. J., and A. P. Skouby, *Acta physiol. Scandinav.*, **18**, 190, 1949.
[4] *Ibid.*, **21**, 286, 1950.
[5] Hensel, H., and Y. Zotterman, *J. Neurophysiol.*, **14**, 425, 1951.

warmth can be stimulated by irradiation with heat rays, measuring the radiant energy necessary to evoke a sensation.[1]

Cold receptors discharge continuously when they are maintained at a constant temperature below a certain level, as is shown by records of spike potentials in the afferent fibers of the cat's tongue. This stationary threshold temperature varies for individual receptors; some may discharge at a low rate at 41°C., but for the majority the threshold lies between 35 and 40°C. The maximum frequency, about 10 impulses per second, is observed between 25 and 35°C.[2] At lower temperatures frequency of discharges diminishes gradually until between 10 and 15°C. a secondary peak two-thirds of the maximum frequency appears. The discharges cease (cold paralysis) when the temperature of the thermode on the tongue falls to 10 to 12°C. If the temperature is raised above 45°C., the cold receptors also discharge, frequency of discharge increasing to seven to eight per second at 50°C. (paradoxical cold).[3]

Warm receptors also discharge when maintained at a constant temperature. Above 47°C. (occasionally 50°C.) and below 20°C. no discharges were observed; maximum frequency (three to four impulses per sec.) occurred at 37 to 40°C.

A sudden fall in temperature causes a rapid increase in the rate of discharge of cold receptors up to a maximum (140 impulses per second) over ten times the frequency of the continuous discharge. The rate of discharge then falls and remains at a constant level dependent on the constant temperature level. If the receptors are rapidly warmed, they cease to discharge; and after a time, if the final temperature is lower than the upper threshold for constant discharge, the impulses reappear. This postexcitatory depression is dependent on the rate of warming and the temperature level; if warming takes place very slowly, the discharge does not cease but continues at a lower level.

A sudden rise in temperature stimulates receptors for warmth; there is an initial rise in the rate of discharge and a fall to a lower constant

[1] Herget, C. M., L. P. Granath, and J. D. Hardy, *Am. J. Physiol.*, **134**, 645; **135**, 20 and 426, 1941.
[2] Hensel, H., and Y. Zotterman, *Acta physiol. Scandinav.*, **23**, 291, 1951.
[3] Dodt, E., and Y. Zotterman, *Acta physiol. Scandinav.*, **26**, 358, 1952.

frequency if the constant temperature level lies between the temperatures at which this is observed. A sudden fall of 8 to 15°C. may cause discharge of a "paradoxical" burst of impulses. Rapid cooling of nerve fibers can stimulate them and evoke spike potentials; this discharge has the same short latency and time course as that of paradoxical warmth,[1] so it is possible that this sensation is caused by stimulation of superficially placed fibers innervating receptors for warmth.

In man changes of outside temperature evoke a sensation if they occur at a certain rate.

The threshold for cold is a fall of 0.004°C. per sec.; for warmth it is a rise of 0.001°C. per sec. In both cases the change must last at least 3 sec. if a very small area is stimulated. Stimulation of a larger area lowers the threshold,[2] probably because a greater number of receptors are stimulated simultaneously and there is summation of their effects. Sensations evoked from receptors situated far from each other can also be summated. When the entire body surface is stimulated, a rise in skin temperature of 0.0008°C. per sec. produces a sensation of warmth.[3] Certain workers make a distinction between warm and hot sensations and determine the threshold for each. Perhaps there are two types of receptor stimulated by an increase in skin temperature, one with a low threshold, which gives the sensation of warmth, the other with a high threshold, which gives the sensation of heat. Temperatures above 45°C. stimulate simultaneously receptors for warmth, cold (paradoxical cold), and pain; the sensation "hot" could therefore be due to a complex pattern of stimulation involving several types of receptors, while stimulation of the receptors for warmth alone by a small increase in skin temperature might give the sensation of warmth. Receptors for warmth soon cease to discharge at temperatures above 45°C.; therefore it has been suggested that the sensation "hot" is initiated by stimulation of warmth receptors and paradoxical stimulation of cold receptors but is maintained by impulses from cold and pain receptors.[4] Observed facts can thus be explained without having to suppose the existence of two types of receptor for warmth and heat, for which there is no morphologic foundation.

[1] BERNHARD, C. G., and R. GRANIT, *J. Gen. Physiol.*, **29**, 257, 1946.

[2] HARDY, J. D., and T. W. OPPEL, *J. Clin. Investigation*, **16**, 533, 1937.

[3] OPPEL, T. W., and J. D. HARDY, *J. Clin. Investigation*, **16**, 525, 1937.

[4] DODT, E., and Y. ZOTTERMAN, *Acta physiol. Scandinav.*, **26**, 345, 1952.

There are considerable regional differences in sensitiveness to changes in temperature. Parts of the skin that are usually covered by clothing, such as the chest and abdomen, are more sensitive; the nose is more sensitive than the cheeks, and the anterior aspect of the forearm is more sensitive than the dorsal one. The mucosa of the mouth has a high threshold; it can be submitted to temperatures (*e.g.*, hot drinks) that are painful to the skin. The conjunctiva, the margin of the cornea, and the genital mucosa are sensitive to cold but not to warmth. Heat applied to these mucosae evokes an unpleasant sensation, which becomes definitely painful if the intensity of the stimulus is increased. The esophagus, stomach, and rectum are sensitive to extreme temperatures (cold or hot), but other viscera and tissues are not sensitive to changes in temperature.

Sensitiveness to temperature changes is modified by several factors. The most important seems to be the condition of the circulation in the skin. If the circulation is stopped, an increase in temperature well below the maximum tolerated by the normal skin becomes intolerable. When the blood is circulating, it carries away heat, and the rise in temperature is thus less than when the circulation is stopped. CO_2 increases sensitiveness to heat. Menthol lowers the threshold of cold receptors and increases that of warm receptors, so that the slightest movement of air, which produces a loss of heat by the skin, gives rise to an intense sensation of cold. Acetylsalicylic acid lowers the threshold for warmth.[1]

Difference thresholds. An increase in the intensity of cold or warm stimulation of the skin produces an increase in the intensity of sensation according to Fechner's law,[2] up to a certain strength of stimulus. Stronger stimuli evoke painful sensations; this occurs when cooling or heating act as noxious stimuli and set free algogenous substances which stimulate pain receptors. With radiant heat the threshold for pain is reached when irradiation is increased to approximately

[1] WICKLER, A., H. GOODELL, and H. G. WOLFF, *J. Pharmacol. & Exper. Therap.*, **83**, 294, 1945.

[2] HARDY and OPPEL, *loc. cit.*

2,000 times the energy required to evoke the threshold sensation of warmth. There are about ninety steps of intensity of sensation distinguishable between the thresholds for warmth and pain.

The difference threshold of intensity is at a minimum when the temperature of the skin is between 27 and 32°C. When the skin is either very cold or very warm, a change of several degrees is needed for temperature differences to be appreciated.

The difference threshold of surface for radiated heat is approximately 3 sq. cm. on the skin of the forehead.

Adaptation. Cold and warmth receptors undergo rapid adaptation, and if the temperatures are not extreme the sensation of cold or warmth soon diminishes.

Contrast phenomena are intense, as is commonly experienced on passing from a warm to a cold environment or vice versa.

The sensory image usually persists for a short time; a cold or hot coin placed on the skin is still felt for a few moments after it has been removed, even though skin temperature begins to rise or fall as soon as the stimulus has ceased to act.

Localization. Cold and warm sensations are referred to the receptor, and there is fairly good localization of the area stimulated. There is very little "projection" into the environment.

The affective tone is usually high; moderate stimuli produce pleasurable sensations. Strong stimuli damage the cells and stimulate pain receptors.

Reactions to cold and heat. (See Chap. 48 and "Hypothalamus," Chap. 84.) Stimulation of the receptors for cold provokes a series of reactions which diminish loss of heat from the body, such as constriction of the arterioles and reduction of blood flow through the skin, erection of cutaneous hair, etc. If cooling is sufficiently marked and prolonged, heat production increases owing to increased muscle tonus and shivering.

Stimulation of the receptors for warmth produces another series of reactions which increase loss of heat, such as cutaneous vasodilatation with an increase in blood flowing through the skin, sweating, etc.

These reactions have a higher threshold than the threshold for the corresponding sensation of cold or warmth.

PAIN

Pain is a conscious state in which there is a more or less high affective tone of unpleasantness, which increases to distress when it becomes more severe, accompanied by reactions that tend to remove or evade the cause that provokes it. Physical pain is produced by alterations in the structural or functional normality of the organism. Sensitiveness to pain is greatly developed in the skin and also exists in other tissues, but in some tissues there are no receptors for pain, and they can be damaged or even destroyed without giving rise to it.

Pain is a specific modality of sensation. At one time it was considered that strong stimulation of any receptor evoked pain. This is not so; the tactile receptors of the skin can be stimulated so as to produce discharges of maximum frequency in the afferent nerve, without any of the reactions typical of pain.[1] A strong stimulus applied to a differentiated receptor can provoke pain if it produces damage and thus acts on specific pain receptors. Certain sensory organs, *e.g.*, Meissner's corpuscles, are innervated not only by a myelinated fiber, which conducts tactile impulses, but also by a fine unmyelinated fiber, which might be the receptor and conductor of pain impulses in response to a strong stimulus.

The specific nature of the sensation of pain has been fully demonstrated by the following observations: (*a*) pain alone, and no other sensation, is evoked on stimulating certain areas of the skin in which there are free nerve endings and no other type of receptor,[2] whereas areas where there are no free nerve endings are insensitive to pain, and a needle can be stuck into the skin without causing pain; (*b*) stimulation of certain structures elicits pain sensations only, *e.g.*, the pulp of the teeth, the arteries of the brain, the middle meningeal and the temporal arteries, etc;[3] (*c*) accidental or surgical transection of certain tracts of the spinal cord suppresses pain without loss of other modalities of sensation (see Chap. 74); (*d*) analgesic drugs may raise the threshold of pain, without modifying other sensory thresholds.[4]

[1] Adrian, E. D., McK. Cattell, and H. Hoagland, *J. Physiol.*, **72**, 377, 1941.

[2] Woollard, H. H., G. Weddell, and J. A. Harpman, *J. Anat.*, **74**, 413, 1940.

[3] Ray, B., and H. G. Wolff, *Arch. Surg.*, **41**, 813, 1940.

[4] Wolff, H. G., J. D. Hardy, and H. Goodell, *J. Clin. Investigation*, **20**, 63, 1941.

Receptors. Pain is provoked by the stimulation of free nerve endings, which have been carefully studied in the cornea (Fig. 390*e*) and in the skin.[1] A cutaneous area is innervated by myelinated nerve fibers approaching it from all directions to form the cutaneous nerve plexus. Fibers in the plexus ramify but do not fuse with one another. Branches from one fiber may innervate many receptors, but these are all of the same type. The fibers for pain emerge from the plexus and give rise to fine ramifications or nerve nets, which emit free nerve endings, often beaded in appearance. These nerve nets are found mainly at two levels; in the deeper layers of the epidermis and in subepidermal tissues near the blood vessels. Some of the nerve nets are connected to myelinated, and others to nonmyelinated, fibers. Differentiated cutaneous receptors, such as Meissner's corpuscles, Krause's end-bulbs, etc., are supplied not only with their "specific" fiber, but also with a fine fiber similar to those of nerve nets.

The area covered by a sensory unit of pain varies in different parts of the body. Thus a single nerve net in the human forearm has been found to be 7.5 mm. in diameter, and one in the thumb pad of the monkey only 1.5 mm. (Weddell). In the cornea one fiber may supply a whole quadrant, *i.e.*, an area approximately 1 cm. in diameter. There is a considerable amount of overlapping of the territories of the different units and of the areas innervated by different cutaneous nerves and dorsal roots.

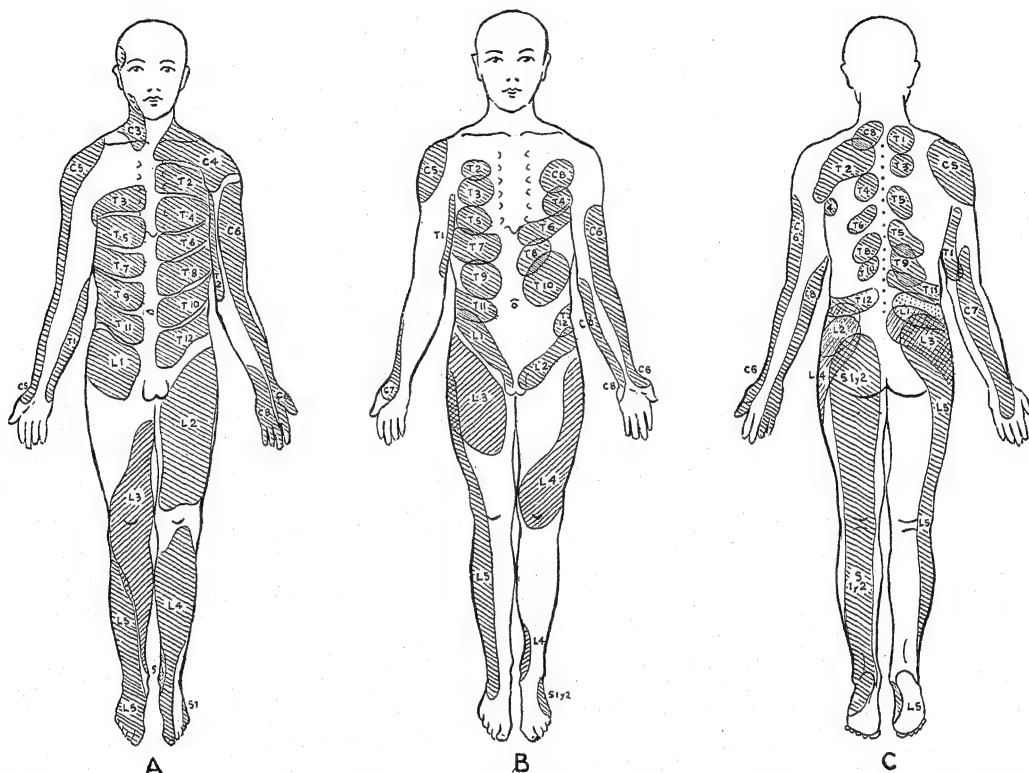



Fig. 392. Areas innervated by sensory spinal roots. *A*, cutaneous areas (dermatomes), according to Foerster; *B* and *C*, deep areas, according to Kellgren. (*Lewis, T., "Pain," Macmillan, New York, 1942.*)

The dorsal spinal roots supply innervation to the skin in overlapping bands, called "dermatomes," which do not coincide with the areas innervated by peripheral nerves, these being made up of fibers from several dorsal roots. Sherrington first delineated the territories of the dermatomes in the cat by severing three roots above and three below the intact root to be studied; a sensitive segment was thus isolated

[1] Weddell, G., *loc. cit.; J. Anat.*, **75**, 346, 1941; **77**, 49, 1942.

between two areas of complete anesthesia. The dermatomes were found to overlap considerably, so that any point in the skin was innervated by two or even three dorsal roots. A similar dorsal-root distribution has been found in the monkey and in other species.

Table 94. Distribution of Sensitiveness to Pain

Sensitive Structures	*Insensitive Structures*
Skin	
Superficial mucosae: conjunctiva, nasal, buccal, anal, genital	
Muscles, tendons, fasciae	
Periosteum, synovial membranes	Bone
Perivascular tissues	Blood vessels
Intracranial structures:	Intracranial structures:
Tissues around arteries and sinuses	Dura
	Pia and arachnoid
Dura near the large meningeal vessels and sinuses	Cerebral parenchyma
	Ependymal canal
Origin of fifth, ninth, and tenth nerves	
Pleura (parietal and diaphragmatic)	Pleura (visceral)
	Lung
Base of fibrous pericardium	Pericardium, visceral and parietal serous membranes
Myocardium (ischemia)	
Subperitoneal tissue (parietal and diaphragmatic)	Esophagus (except to distention)
Mesentery (perivascular tissues)	Peritoneum, omentum
	Liver, spleen, kidneys
Pancreas, duodenum, hepatic pediculum	Stomach, intestine, colon, rectum*
Ureter, renal pelvis	Bladder
Periovarial tissues	Ovary, uterus, vagina
Peritesticular tissues	Testes

* Distention and contractures provoke pain.

Observations made in cases of herpes zoster (a virus infection of the posterior-root ganglia, commonly known as "shingles") and in patients in whom dorsal-root section was necessary for therapeutic reasons[1] have shown that in man also there is a dermatomal distribution. The root segments can be fairly well distinguished in the trunk (Fig. 392*A*). They slope downward and spread out from the dorsal to the ventral surface. In the limbs, especially in the lower limb, the dermatomes are not so clearly discernible, because several spinal segments concur in each limb, but there is also a dermatomal distribution of the nerve fibers.

[1] Foerster, O., *Brain*, 1, 56, 1933.

The dermatomal area of pain is smaller than that of temperature, the largest being that of touch. Dermatomal charts are by no means accurate; nevertheless, they give a general idea of dorsal-root distribution.

Cutaneous dermatomes do not coincide with the distribution of the dorsal root in the deeper structures (Fig. 392*A*, *B*, and *C*). Kellgren[1] has studied this distribution by injecting 0.3 cc. of a 5 per cent NaCl solution in the interspinous ligament, a procedure that elicits short-lasting pain referred to the territory innervated by the corresponding root.

Distribution of sensibility to pain. Pain sensitiveness is most developed in the skin but is not limited to it; other structures, listed in Table 94, are also sensitive to pain. Trauma evokes little or no pain in somatic muscles, but when they contract in conditions of ischemia a sustained pain is felt, the intensity of which is related to the amount of work performed by the muscle. The heart can be handled and cut without pain, but ischemia due to occlusion of a coronary artery produces the intense pain of angina pectoris. The digestive tract is also insensitive to trauma and can be cut and burned without provoking any sensation, but distention of the esophagus or any part of the gastrointestinal tract and strong contracture of the stomach or the intestine (colic) can be very painful.

Sensory threshold. Pain can be provoked by many different agents; mechanical, thermal, electrical, chemical, etc. The receptors are apparently anelective. "But it is to be remarked that these agents, regarded as excitants of skin-pain, have all a certain character in common, namely this, that they become adequate as excitants of pain when they are of such intensity as threatens damage . . . and they are all able to excite when applied to naked nerve directly."[2] Free nerve endings are an adequate mechanism for this type of excitation, in which there must be a response to a great variety of stimuli. A differentiated receptor, with specific irritability, would be less well adapted to such an end.

Pain can be investigated by recording either the sensation of pain or reactions to painful stimuli, such as a defense reflex, a muscle twitch, etc. In the first case the subject gives a verbal

[1] Kellgren, J. H., *Clin. Sc.*, 4, 35, 1939.

[2] Sherrington, C. S., "The Integrative Action of the Nervous System," Yale University Press, New Haven, 1926, p. 227.

report of his sensations, and unless special precautions are taken, subjective differences in appreciation may have great influence on the results. The subject should be instructed in the procedure to be used for the determination of the threshold. He should be in a condition of objective attention and free from fear of harm or great suffering. Apprehension may alter the results to a considerable degree, and voluntary or involuntary inattention, such as is produced by an intercurrent stimulus, raises the threshold.

"The pain threshold sensation may be defined as the lowest perceptible intensity of pain. The pain threshold stimulus is the amount of stimulus required to induce threshold pain,"[1] and it is most conveniently stated in terms of physical energy, *e.g.*, the force of a prick, voltage of an electric current, concentration of a caustic substance, radiant energy of heat, etc. A method commonly in use is to prick the skin with varying strengths; there are special appliances (algometers) that permit fine graduation of the force used. This method deforms the skin and therefore may stimulate touch receptors at the same time.

Bishop[2] has used high-voltage and low-current sparks to stimulate the receptors without deforming the skin. He has thus been able to excite separately touch and light-pressure corpuscles, the receptors attached to hair follicles, and cold receptors, but not the more deeply placed warmth receptors. Pain, with this method of stimulation, has a lower threshold than touch, except on the finger tips. With weak stimuli a painless prick, surrounded by an aura of itching, is felt after a relatively long latent period; this becomes a prickling pain when the strength of the stimulus is increased.

Radiant heat has been used to stimulate pain receptors.[3] The results are expressed in gram-calories per square centimeter per second. With weak stimuli a sensation of warmth is felt, which increases with the strength of the stimulus and sharply changes to one of pricking pain when the pain threshold is reached. This threshold is about 1,000 times the radiant energy needed to produce the minimum sensation of warmth.

Bishop[4] points out that three thresholds may be recognized when pain endings are stimulated: (*a*) a nondescript threshold contact sensation; (*b*) a sharp prick, perceived as painful, but without emotional protest; (*c*) reaction to pain manifested by defense reflexes and emotional protest.

The pain threshold is fairly constant in the same individual if there are no disturbing factors. The pricking-pain threshold provoked by radiant heat applied to the forehead remains constant within ± 3 per cent on successive determinations in the same individual. Daily variations were found to be within ± 8 per cent, and sometimes the same threshold was maintained for several weeks. Individual differences were not great (± 15 per cent). Young people seem to be more sensitive to pain than the aged. There is a significant correlation between the threshold of skin pain provoked by radiant heat and that of visceral pain provoked by distention of the esophagus; persons highly sensitive to one kind of pain are also highly sensitive to the other.

Factors that cause variations in pain threshold. Psychic factors are of great importance in pain and may exert considerable influence on the pain threshold.

Autosuggestion and external suggestion, such as is produced by giving a substance with supposed analgesic properties (placebo) and by hypnosis, may increase the threshold sufficiently to make painless the action of agents that destroy the tissues. Suggestion and autosuggestion can also lower the threshold; thus the lowest thresholds observed have been found in cases of hysteria. One of Hippocrates' aphorisms states that, if two pains occur together in different parts of the body, the stronger weakens the lesser. In persons suffering spontaneous or previously provoked continuous pain, the threshold for cutaneous pain is raised.

Damage to the skin that diminishes the thickness of the epidermis, sunburns, inflammatory processes, etc., lower the pain threshold. On the other hand, thickening of the epidermis (callosities, etc.) raises it.

Analgesic drugs raise the pain threshold. If sufficiently intense pain is experienced early in the course of the action of an analgesic agent, the pain threshold is not raised much above the normal, but the reaction to pain is considerably diminished, *i.e.*, the subjects perceive pain but remain indifferent to it.

[1] Wolff, H. G., and J. D. Hardy, *Physiol. Rev.*, **27**, 167, 1947.

[2] Bishop, G. H., *J. Neurophysiol.*, **6**, 361, 1943.

[3] Hardy, J. D., H. G. Wolff, and H. Goodell, *J. Clin. Investigation*, **19**, 649, 1940.

[4] Bishop, G. H., *Physiol. Rev.*, *loc. cit.*

There is no peripheral spatial summation of pain. The threshold does not vary with the surface of skin irradiated; after the minimum area has been exposed, the threshold is not reduced, nor does the sensation increase on irradiation of a larger surface, but the discomfort is greater as the irradiated surface increases.

Approximately twice the threshold energy for pain is needed to cause tissue destruction; the organism is thus warned of the imminence of damage. Wolff and Hardy point out that "there is no need in the body economy for a wide range of experience between the onset of pain and tissue damage since with the first pain impulses the organism receives warning that the limit of safety has already been passed."[1]

Hyperalgesia. Two different phenomena are included under this term: (*a*) lowering of the threshold; (*b*) an increase in the intensity of pain sensation. They can occur simultaneously or separately.

Mechanical trauma, burning, freezing, irritating substances, ultraviolet rays, etc., produce, after a period ranging from a few seconds to several hours, a triple response: (*a*) capillary dilatation; (*b*) arteriolar dilatation; (*c*) local edema, accompanied by spontaneous pain and by hypersensitiveness (low threshold and increase in response) to pain-producing agents (friction, cold, warmth, etc.). The area of redness and hyperalgesia extends farther than the damaged tissue, especially when the injury is severe. When no visible lesion has been produced, the condition is known as "erythralgia," because of the redness caused by vasodilatation of the hypersensitive area.[2] If there is no spontaneous pain, it is easily awakened by increasing the tension of the skin, *e.g.*, by stretching it with the fingers, or warming it up to 32 to 34°C., or hindering the circulation. Rubbing immediately produces intense pain of short duration, followed after an interval of 15 sec. by another pain which increases gradually and lasts longer. If the circulation is stopped, the second pain is more severe and prolonged. Apparently it is produced by algogenous substances set free by friction (see "Mechanism of excitation," further on).

These facts are of importance in the treatment of cutaneous pain. The patient should be placed so as to avoid tension of the damaged skin. Excessive warmth must also be avoided, so as to keep the skin below the critical temperature. All contacts and friction should be carefully prevented. Immobilization in a plaster cast, leaving the injured area uncovered, has been used successfully in the treatment of pain caused by burns.[1]

[1] Wolff and Hardy, *loc. cit.*
[2] Lewis, T., *Clin. Sc.*, **1**, 175, 1933–1934.

Discrimination in intensity of pain. Subjective intensity of pain varies considerably from one person to another. There are no data to support the idea that this variation is due to anatomic or physiologic individual differences such as a greater or lesser number of receptors, differences in the resistance to agents that set free algogenous substances, variations in these substances, etc. Injury of the same type and severity produces many and loud complaints from some individuals and is well tolerated by others; the psychic factors are of predominant importance.

By means of the radiant-heat method of provoking pain it is possible to demonstrate the capacity to discriminate the intensity of pain in relation to strength of stimulus.[2] Applying stimuli ranging in intensity from the threshold (218 μcal per sec. per sq. cm.) to 45 per cent above (480 μcal per sec. per sq. cm.), 21 just-noticeable differences of pain intensity can be distinguished. Differentiation of intensity is not possible if stronger stimuli are applied.

Adaptation. There is a certain degree of adaptation in pain receptors, but much less than in touch and thermal receptors. The *duration* of pain varies considerably; it is often of great use in the clinical diagnosis of the cause of pain. Figure 393 shows diagrammatically the time-intensity curve of several types of pain.

Quality of pain. The quality of pain can be expressed only vaguely and in terms of the cause that most frequently produces it. Two types of cutaneous pain have been distinguished since Gad and Goldscheider first pointed them out in 1892. One type of pain has a pricking quality; it begins and ends abruptly, it is fairly well localized, and there is good discrimination in the degrees of intensity. The second type has a burning quality; it is slow in onset, recedes more slowly, and is more diffuse and less well localized than pricking pain. By means of the thermal-radiation method the threshold of "burning" pain was found to be 13 per cent lower than the threshold for "pricking" pain.[3] Pricking pain

[1] Zeno, L., *Rev. Asoc. méd. argent.*, **57**, 854, 1943.
[2] Wolff and Hardy, *loc. cit.* (quotation from Hardy, Goodell, and Wolff).
[3] Bigelow, N., J. Harrison, H. Goodell, and H. G. Wolff, *J. Clin. Investigation*, **16**, 525, 1937.

is felt more rapidly ("first" or "quick" pain) than burning pain ("second" or "slow" pain). The impulses of the former are conducted by fibers at the rate 10 to 90 m. per sec., and those of the latter by slow-conducting C fibers (see Chap. 74). The second type of pain is similar to deep pain arising in the viscera and deep-lying structures, such as the joints and muscles (see Chap. 73).

NEEDLE PRICK
PULSATING TOOTHACHE
INTESTINAL COLIC
ANGINA
CONTINUOUS PAIN, FLUCTUATING

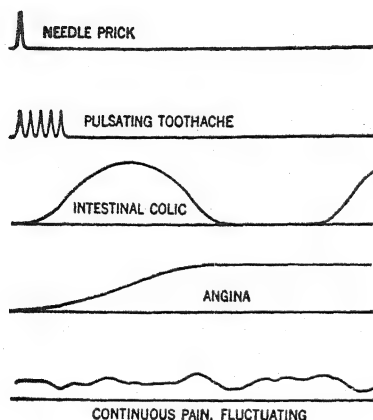



FIG. 393. Diagrams illustrating time-intensity curves of pain. (*Lewis, T., "Pain," The Macmillan Company, New York, 1942.*)

The conjunctiva, cornea, and nasal and genital mucosae have a pain sensibility of the second (burning) type. These structures are hypersensitive to pain. They also have cold receptors, but they are insensitive to warmth.

Section of the descending root of the trigeminal nerve produces analgesia in the territory it innervates, but has little effect on sensitiveness to touch. In the cornea mechanical stimuli are no longer painful, but arouse a tactile sensation. Therefore it is possible that, in normal subjects, hypersensitiveness to pain masks the sensation of touch. If this is so, free nerve endings would act as touch receptors, since in the central part of the cornea there are no other nerve structures, and in the periphery there are only free nerve endings and Krause's end-bulbs.

Localization. Pricking pain in the skin and the mucosae of the mouth and anus is well localized. The error is less than 1 cm. on the hand. Accuracy in localization is not due to the simultaneous touch sensation, as it persists after touch has been suppressed by ischemia. Localization is less precise in mucosae with pain sensibility.

Pain reaction. Perception of pain is often, but not always, accompanied by reaction to pain, which has psychic, somatic, and visceral components. The emotional state evoked by pain consists in aversion to, and a protest against, pain; an intense desire to avoid or suppress the cause of pain; and fear. Somatic reactions to pain are of two types: (*a*) increase of muscular activity, manifested by defense reflexes and the "fight or flight" pattern of behavior, often accompanied by vocal protest; (*b*) depression of muscular activity (quiescence) together with localized contractures, immobilizing the painful parts of the body. The first type of somatic reaction is associated with increased activity of the sympathicoadrenal system, *i.e.*, tachycardia, cutaneous and abdominal vasoconstriction, hypertension, mydriasis, hyperglycemia, etc. Visceral reactions characteristic of the second type are bradycardia, hypotension, nausea, and vomiting ("sickening" pain). Sweating is common to both. The first type of reaction is usually observed in response to pain evoked from the skin; the second is associated with deep somatic or visceral pain. Painful stimulation may cause a decrease in urinary output by two different mechanisms: (*a*) renal vasoconstriction, evoked reflexly or by a humoral mechanism (sympathin, adrenaline), since it persists after denervation of the kidney;[1] (*b*) increased secretion of posterior hypophyseal antidiuretic hormone.[2]

Perception of, and reaction to, pain are often dissociated. Pain may be perceived without causing a reaction, *e.g.*, during combat, games, sexual excitement, or mystic experience, or in patients suffering from traumatic shock, certain lesions in the brain (*e.g.*, after frontal lobotomy), or psychic disturbances. Many of the components in the reaction to pain can be voluntarily suppressed, *e.g.*, vocal protest, and some of the somatic defense reflexes. In this case visceral reactions persist and may produce syncope. Certain drugs, such as ethyl alcohol and morphine, depress the reaction to pain. Wolff and his associates[3] have analyzed bodily reactions to pain by measuring changes in the electrical

[1] BROD, J., and J. H. SIROTA, *Am. J. Physiol.*, **157**, 31, 1949.

[2] KELSALL, A. R., *J. Physiol.*, **109**, 150, 1949.

[3] WOLFF, H. G., J. D. HARDY, and H. GOODELL, *J. Pharmacol. & Exper. Therap.*, **75**, 38, 1942; CHAPMAN, W. P., and C. M. JONES, *J. Clin. Investigation*, **23**, 81, 1944.

resistance of the skin on the palm of the hand, and by observing wincing, as revealed by narrowing of the eyelids. These "alarm" reactions are much more variable than the pain-perception threshold. Drugs, psychic conditions, etc., may increase or lower the threshold of reaction, without significantly altering the pain-perception threshold.

Mechanism of excitation. Free nerve endings, *i.e.*, pain receptors, are stimulated directly by the many noxious agents (mechanical, chemical, etc.) that cause pain.

Tension is a frequent cause of pain; *e.g.*, when a hair is pulled the pain is due to tension exerted on the nerve endings, pulsating pain is produced by distention of the tissues at each systolic wave, and the immediate relief felt on opening an abscess is due to release of tension.

The free nerve endings sensitive to pain are also stimulated by a chemical factor released when the cells are injured. The following facts support this statement:

1. When a muscle contracts after its circulation has been stopped, pain develops and increases in intensity until movement is no longer possible. The intensity of pain depends on the amount of work done, and if contraction ceases it does not increase, but persists at the same level as long as ischemia is maintained. It disappears in 3 to 4 sec. after the circulation has been reestablished. Contraction produces an "algogenous" substance, which diffuses into the interfibrillar spaces and excites the free nerve endings.[1] In normal conditions this substance is removed as soon as it is produced and does not reach the concentration needed for stimulation of nerve endings. Intense and prolonged exercise leads to accumulation of the substance and thus to lowering of the pain threshold, and the muscles feel sore. Anoxia alone is not the cause of pain; there must also be contraction of the muscle. Anoxia plays a necessary part, however, as is shown by the fact that contraction becomes painful if the muscles are supplied with anemic blood (lack of oxygen carriers) or with blood that is poor in oxygen as a result of breathing air with a low oxygen partial pressure. In this case there is no vascular spasm; on the contrary, asphyxia causes vasodilatation. Nevertheless, vascular spasm can produce pain on contraction of the muscles, because it reduces the blood supply. Obstruction of a blood vessel has the same effect. Angina pectoris is caused by thrombosis or spasm of a coronary artery; the heart continues to beat (contract) in spite of the deficient circulation, and an intense pain is felt which persists until the circulation improves.
2. Mechanical stimulation (rubbing, hitting) of an area of hyperalgesia immediately provokes pain of short duration, followed 15 sec. later by a second pain of longer duration.[1] If the circulation is stopped the second pain persists. Ischemia alone can provoke it, and it disappears when the circulation is restored.

[1] Lewis, T., G. W. Pickering, and O. Rotschild, *Heart*, **15**, 359, 1929; **16**, 1, 1931.

The *chemical nature* of this algogenous substance (Lewis's P factor), which can be extracted from a piece of traumatized skin, is yet unknown. Perhaps there is more than one such substance. It is not a protein, nor is it acetylcholine, histamine (which causes itching), or lactic acid.

Effects of denervating the skin. Section of a cutaneous nerve, if the severed trunk is sufficiently large, produces complete anesthesia over a certain limited area, surrounded by an area of hypoesthesia involving all types of sensation, in which sensitiveness decreases progressively from the normal skin to the central area of complete anesthesia. In the intermediate zone the sensory units of the cut nerve have been destroyed, but those of other nerves still remain and are responsible for the remnant of sensibility. When the nerve regenerates, sensitiveness increases gradually. Touch and pain are recovered first, together with cold; warmth is recovered only later. The threshold is higher, but the response (intensity of sensation) is greater and is referred to the periphery, *i.e.*, to the site where the receptor is normally situated. These abnormal qualities in sensation are possibly due to stimulation of naked fibers growing out into the denervated area.[2]

Head and Rivers,[3] analyzing experiments of cutaneous denervation performed principally on Head himself, distinguished two types of sensibility.

[1] This second pain, due to liberation of a chemical factor, should not be mistaken for the second pain due to slow-conducting fibers.

[2] Trotter, W., and H. M. Davies, *J. Physiol.*, **38**, 134, 1909.

[3] Head, H., and W. H. R. Rivers, *Brain*, **31**, 323, 1908; Head, H., "Studies in Neurology," vol. 1, London, 1920.

One was considered phylogenetically old, and included pain and the extremes of temperature (below 20 and above 37°C.); they called it "protopathic" sensation. The other, more recently developed, permits fine discrimination in tactile and thermic sensations; they called it "epicritic" sensation. According to Head, protopathic sensation is integrated in the thalamus, and epicritic sensation in the cerebral cortex. Normally the former is overshadowed by the latter, but denervation suppresses epicritic sensibility in the intermediate zone, where only protopathic sensibility persists; during the period of regeneration this latter type is recovered first. Head's observations, based on a rather small number of cases, have been confirmed only in part. Confirmation is lacking for some fundamental aspects needed to support his theory of the two sensibilities, which is not generally accepted.

Congenital analgesia. Pain, in spite of its importance as a defense mechanism and in the building up of the mental world, is not indispensable for survival or for an apparently normal physical and mental life. Several cases of congenital absence of sensitiveness to pain have been described.[1] These subjects did not feel pain when stimuli such as pricking and burning were applied to the skin, when the esophagus was distended, when muscles were made to contract in conditions of ischemia, or when the testes were compressed (although this provoked nausea), nor did the injection of histamine provoke headache (see Headache, Chap. 73). Noxious stimuli did not provoke the usual psychic and physical reactions that accompany pain; the subjects did not complain, the majority of defense reflexes were absent, and the psychogalvanic reflex (increase of sweat secretion) brought about by emotional stress was also missing. Some of the simpler defense mechanisms persisted, such as the corneal reflex, tear secretion on irritation of the nasal mucosae, and the pressor reflexes to cold. Tactile, temperature, and proprioceptive sensations were normal, but there was no sensation of itching. These subjects behaved in respect to pain as if they were under light ether or chloroform anesthesia. The pain mechanism seemed to be disconnected from the cortex, because the reactions that required cortical integration were missing. There are no anatomic data that permit localization of the disturbance in these cases.

[1] FORD, F. R., and L. WILKINS, *Bull. Johns Hopkins Hosp.*, **62**, 448, 1938; KUNKLE, E. D., and W. P. CHAPMAN, *Proc. Assoc. Res. Nerv. Ment. Dis.*, **23**, 100, 1943.

Itch.[1] This is a disagreeable sensation that induces scratching of the part where it is felt. It is originated exclusively in the epidermis and cannot be provoked where the epidermal layers have been removed. Less well localized than touch and pain, it has a tendency to irradiate and to persist after the originating stimulus has ceased to act, and its degrees of intensity are not well differentiated.

Itching is associated with the sensation of pain:

1. It has the same features as deep pain (diffuseness, poor localization, and poor discrimination in the degrees of intensity).
2. Agents that provoke a triple response also produce itching. If they are weaker, a tickling sensation is felt; if they are stronger they are painful.
3. When there is dissociation of the different modalities (due to disease, local anesthesia, asphyxia of the nerve, or any other cause), pain and itching disappear or are reestablished at the same time.
4. When there is complete analgesia, itching is not felt. Loss of touch without analgesia does not entail the loss of itch.
5. Itch varies independently of the temperature senses.
6. Bishop[2] has provoked itching by stimulating pain fibers without deformation of the skin, by means of high-frequency low-intensity stimuli. Stronger stimuli provoked a prick sensation; if they were repeated at adequate intervals, they were followed by a sensation of itch.
7. Stimuli that provoke tickling, itching, and the second pain produce nerve impulses in slow-conducting fibers (Erlanger and Gasser's C fibers).[3]
8. Nerve impulses provoked by itching are conducted in the spinothalamic tract together with the pain impulses. Itch and pain in cases of anal or vaginal pruritis are suppressed by section of these tracts.

Itching is apparently due to the liberation of a substance similar to histamine (H substance). Intracutaneous injection of histamine produces itching only, if care is taken not to damage the

[1] ROTHMAN, S., *Physiol. Rev.*, **21**, 357, 1941.
[2] BISHOP, G. H., *J. Neurophysiol.*, **6**, 1943, *loc. cit.*
[3] ZOTTERMAN, J., *J. Physiol.*, **95**, 1, 1939.

tissues. Lewis[1] maintains that the chemical agent responsible for itch (H factor) differs from that of pain (P factor). He does not agree with von Frey and Rothman that tickling, itching, and pain are simply increasing degrees of intensity of the same sensation. Pain and itch can be separated from each other when they exist simultaneously by immersion of the skin in water at 40 to 41°C.; itching promptly ceases, but pain is intensified.

Incompletely denervated skin areas are the site of intense itching, which produces violent scratching, leading to injury of the skin. If one or more dorsal roots in the cervical region in cats are severed, ulceration and loss of hair due to scratching are provoked in the borderline areas, which are incompletely denervated so that there are intact fibers from the trigeminal nerve or the following uncut dorsal root. When the head and neck are protected by a leather helmet there are no skin lesions, but as soon as the helmet is removed, the animal scratches violently and the observer can see the skin lesions being produced.[2]

Cortical localization of itch is probably the same as that of pain (parietal lobe). Psychic factors have great influence on itching, both in provoking it (phobias) and in suppressing it.

Adrenaline and ephedrine quickly suppress the itching of urticaria, probably by their effect on the local process that sets free the H factor. Bromides and barbiturates diminish itching by acting on the cortex and the subcortical nuclei, respectively. Bile salts and morphine can provoke intense itching. Morphine acts on the nerve centers, probably on the medulla, below the acoustic nucleus.

Tickling. There are two kinds of tickling. One is produced by deep pressure repeated at adequate intervals in certain regions (armpits, thighs, etc.). The other, known as "skin tickle" is provoked by applying an intermittent light stimulus to the skin, or successive stimulation of adjacent spots at a low intensity. Skin tickle has been attributed to simultaneous excitation of touch and pain receptors,[1] because tickling cannot be provoked (*a*) after pain has been suppressed by section of the spinothalamic tract or of the descending root of the fifth (trigeminal) nerve, operations that suppress pain but do not disturb touch sensations; (*b*) after touch sensations have been suppressed although sensitiveness to pain persists; *e.g.*, when the circulation to a limb is arrested, touch is lost first and pain only later, but tickling is lost together with touch.

The prickling sensation felt when the circulation is stopped apparently arises in receptors situated deeply under the skin.

[1] LEWIS, T., and W. HESS, *Clin. Sc.*, **1**, 39, 1933–1934.

[2] AUBRUN, E. A., *Rev. Soc. argent. de biol.*, **8**, 59, 73, 88, 233, 333, 395, and 482, 1932.

BIBLIOGRAPHY

See page 902.

[1] ZOTTERMAN, J., *loc. cit.*

CHAPTER 73

Visceral and Deep Somatic Sensations

SENSATIONS IN THE deep tissues may be classified as visceral or somatic, according to the structures in which they arise. Visceral sensibility is not so well developed as somatic sensibility; its impulses seldom enter the field of consciousness, and when they do, the sensations are not well localized. There is, however, no essential difference between the two types of sensibility; the same laws are valid and the structural basis is the same for both. The myelinated and unmyelinated sensory fibers of the viscera arise in neurons situated in the spinal and cranial sensory ganglia.

The visceral and somatic systems are closely associated functionally. Stimulation of visceral afferent fibers provokes reactions in (*a*) viscera (viscerovisceral reactions); (*b*) muscles (visceromotor reactions); (*c*) superficial and deep parts of the body's frame (viscerosensitive reactions). On the other hand, stimulation of somatic afferent fibers may provoke responses from the viscera (somatovisceral reactions). Visceral functions are regulated by these reactions. The heart rate, the blood pressure, the secretions and motility of the digestive tract, etc., are controlled partly by reflexes (considered in detail in the respective chapters). Sensory impulses initiating these reflexes travel in fibers that, with few exceptions, form part of the visceral nerves. Usually these impulses do not awaken conscious sensations, except in particular cases, such as hunger and thirst, evacuation of the rectum and bladder, genital reflexes, etc., in which the complete response is integrated by reflex and voluntary somatic movements.

Lesions in the viscera and disturbances in visceral functions in certain cases provoke pain (visceral pain). Impulses thus originated travel to the centers, in afferent fibers often included in sympathetic, and in a few cases in parasympathetic, nerves. Painful impulses arising in the internal aspect of the body wall, in muscle and other somatic structures, will be considered here together with visceral pain.

There are also deeply placed somatic receptors—in connection with the skeletal muscles, tendons, and joints—sensitive to changes in tension and pressure. They form part of the proprioceptive system which regulates the activity of somatic muscles.

VISCERAL AND DEEP SOMATIC PAIN

It is a well-known fact that the parenchyma of the viscera, even of the brain, is not sensitive to handling. In the course of surgical operations made under local anesthesia, the viscera of conscious subjects have been not only handled but also cut and cauterized without arousing any sensation. The viscera, however, are not completely insensitive; the membranes that cover them, the supporting tissue, neurovascular pedicles, and visceral muscles may give rise to localized or irradiated pain, which is often referred to parts of the body distant from the site of the disturbance. This pain is accompanied by reflex visceral and somatic reactions, and by cutaneous hyperalgesia and deep tenderness in definite areas, which vary according to the organ from which the pain arises. Visceral pain is provoked by (*a*) sudden distention, due to congestion or to the accumulation of fluid or gas in hollow viscera; (*b*) traction on the pedicles; (*c*) muscular contraction in conditions of ischemia; (*d*) irritation by chemical agents (algogenous substances) produced by inflammatory processes.

Characteristics of visceral and deep somatic pain. There is no essential difference between pain arising from the viscera and that arising from the deep somatic structures. Both

have an aching quality, both are more or less diffuse and poorly localized, and in both the phenomenon of referred pain can be observed. Pain having all the characteristics, including the reactions, of visceral pain can be aroused by stimulating somatic tissues. Thus, Lewis and Kellgren[1] reproduced the pain of angina pectoris by injecting 0.5 cc. of hypertonic (5 per cent) NaCl solution into the first thoracic interspinous ligament on the left side. Subjects who had suffered from anginal pain could not tell the difference, except for the local pain provoked at the site of the injection. The pain of renal colic was reproduced by injecting hypertonic saline into the first lumbar interspinous ligament, or into the abdominal oblique muscle at the level of the first lumbar dermatome. Pain identical to that of intestinal colic was provoked by injecting hypertonic saline into the rectus muscle of the abdomen.

Localization. In some cases pain is felt in the organ in which it arises (splanchnic pain); *e.g.*, the hepatic pain of a congested liver, deep retrosternal pain of angina pectoris, toothache, earache, etc. The pain, however, is more or less diffuse and localization is always poor.

Deep somatic pain is fairly well localized when it arises from tissues near the body surface, *e.g.*, superficial fascia or tendons, periosteum of bones under the skin such as the tibia, walls of the body cavities (thorax and abdomen), etc. When it arises in deeply placed structures, such as the belly of a muscle, it is diffuse and is often referred to a distant area.

Two factors are of importance for localization of sensory phenomena: (*a*) the extent of cortical representation; (*b*) previous experience of the sensation. The cortical representation of viscera and deep somatic structures is very poor compared with that of the cutaneous surface or of the great cephalic receptors. Moreover, impulses arising in the viscera seldom attain the level of consciousness; there is therefore very little experience of visceral sensation, while exteroceptive impulses are continually awakening sensations. These facts give a reasonable explanation for the poor localization of visceral and deep somatic pain.

Referred pain. Pain localized in a region distant from the site of its origin is known as "referred pain." It is commonly observed in all forms of deep pain, both visceral and somatic. The pain is not well localized but is diffuse and referred to different parts of the dermatome in which it arises; in some cases, when it is particularly severe or prolonged, it may spread to the whole dermatome or even to neighboring dermatomes.

The areas of the deep dermatomes do not coincide with those of the skin dermatomes (see Chap. 72, Fig. 392). Kellgren has determined their distribution by injecting hypertonic saline into the interspinous ligament. Deep-pain areas are concentrated near the mid-line on the ventral and dorsal aspects of the body, and those of the lower cervical, upper thoracic, lumbar, and sacral roots extend far into the corresponding limb. Irradiation and sites of reference of the principal forms of visceral and parietal pain will be considered later.

The mechanism responsible for referred pain and the reactions of deep pain has been the subject of much discussion. Sturge, in 1883, attributed cutaneous hyperalgesia of angina pectoris to increased irritability of the spinal centers caused by the abnormal impulses arising in the heart. Later Head and MacKenzie[1] developed this conception of the "irritable focus" in order to explain referred pain and the reactions of pain. According to MacKenzie, impulses from the viscera normally do not arrive at the cortex but remain at the spinal level. Abnormal impulses create an "irritable focus" so that normal impulses coming from the skin and other structures are "facilitated," *i.e.*, there is summation of abnormal visceral with normal somatic impulses which thus reach the cortex. The sensation aroused is referred to the corresponding cutaneous area, from which impulses habitually come, and not to the viscera, from which impulses are seldom sent. Increased focal irritability would also account for the reactions of pain, such as muscular rigidity, vasomotor and secretory responses, etc. Normal impulses from the muscles and viscera would summate with the abnormal visceral impulses and the corresponding reflexes would be facilitated (Fig. 394).

Hyperalgesia and muscular contracture often persist after the pain has ceased. This fact and certain observations on cutaneous hyperalgesia led Lewis to doubt the existence of the "irritable focus." According to Lewis, hyperalgesia

[1] KELLGREN, J. H., *Clin. Sc.*, **3**, 182, 1937–1938; LEWIS, T., and J. H. KELLGREN, *Clin. Sc.*, **4**, 47, 1939.

[1] MACKENZIE, "Symptoms and Their Interpretation," 2d ed., Shaw and Sons, London, 1912.

is due to liberation of a chemical agent in the sensitive area, which lowers the threshold of pain receptors. The origin of the impulses that release this substance is unknown. Lewis has postulated the existence of special "nocifensor" fibers in the posterior roots, but there is no anatomic evidence to support this hypothesis. Another possibility is the discharge of "antidromic" impulses in the afferent fibers of the dorsal roots (see Chap. 69), but they have as yet not been demonstrated.

Reflex reactions to pain may, in certain cases, be an extra cause of pain; thus, prolonged muscular contraction causes soreness and deep tenderness in the affected muscles. Peripheral changes in areas situated far from the original site of pain can, however, only prolong or increase pain; they cannot account for the initial sensations and reactions. Summation of abnormal visceral impulses with the normal stimulation of the skin is probably the main cause of hyperalgesia.

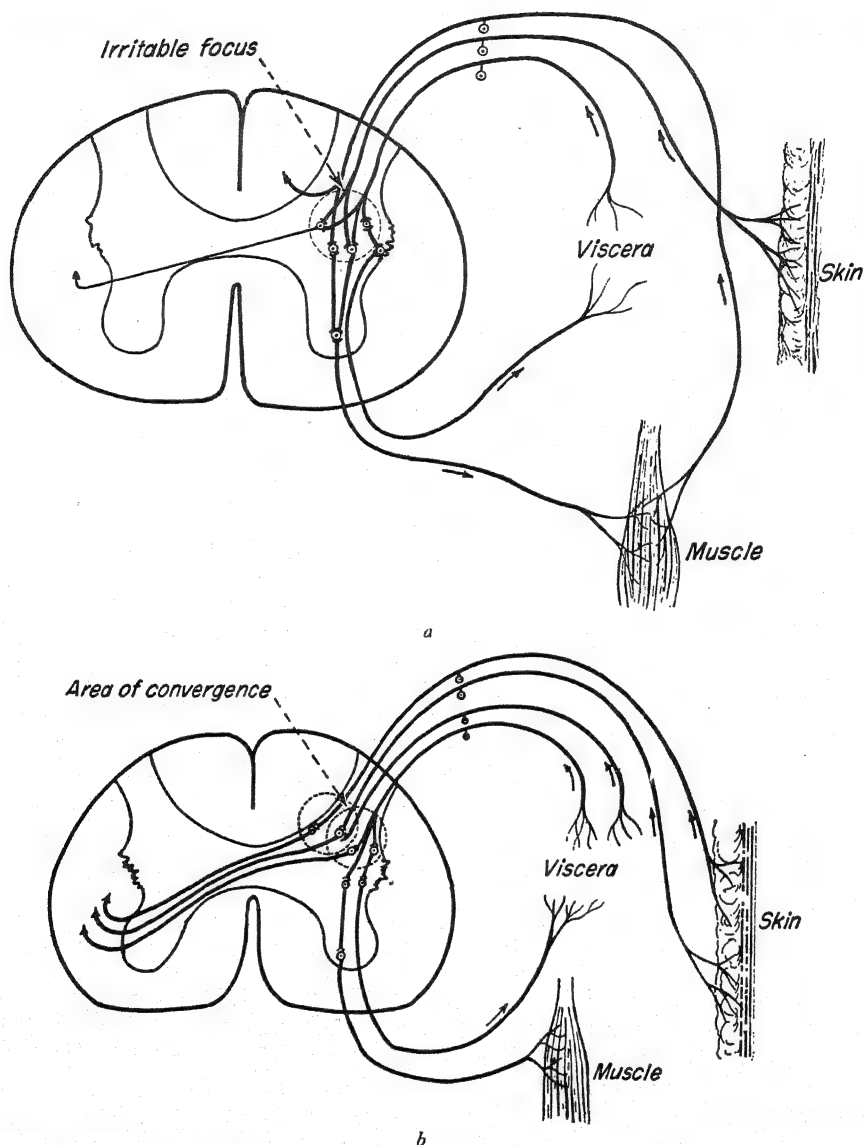



FIG. 394. Diagrams illustrating interpretations of visceral reflexes: *a*, Sturge's and MacKenzie's theory of the "irritable focus"; *b*, the "convergence-projection theory."

An adequate explanation of referred pain is given by the "convergence-projection theory" (Ruch). There is a certain degree of overlapping of the spinal endings of cutaneous and visceral fibers; therefore, some spinothalamic neurons receive impulses from the skin and the viscera. These impulses are transmitted to cortical centers and the sensation aroused is interpreted in the light of the most frequent experience, as coming from the corresponding cutaneous area. Splanchnic pain, *i.e.*, pain localized in the organ in which it arises, is caused by impulses traveling along "visceral" spinothalamic fibers (Fig. 394).

Reaction to visceral and deep pain. Cutaneous pain provokes reactions which form part of the "fight or flight" pattern of response. Deep pain, especially visceral pain, is accompanied by (*a*) cutaneous hyperalgesia; (*b*) tenderness in deep tissues; (*c*) sustained contraction, which immobilizes the painful parts (defense reflexes); (*d*) cutaneous vasoconstriction (paleness) and other vasomotor reactions, which may produce hypertension if the pain is not severe but which cause hypotension and collapse (syncope) if the pain is intense or prolonged; (*e*) nausea and vomiting (sickening pain); (*f*) sweat and other secretions.

Surgical treatment of visceral pain. Visceral pain, when it is severe and prolonged, can be treated by surgical methods, because its nerve paths can be cut without provoking serious disturbance in visceral function. The following methods are used: (*a*) denervation of the organ or region by cutting the nerves in the vascular pedicle, in which case the arteries must be stripped of periarterial tissue; (*b*) extirpation of the sympathetic chain in the region where the pain fibers of the diseased area enter into this chain (*e.g.*, the stellate and upper thoracic ganglia for the heart), or extirpation of the parasympathetic, where the pain fibers are included in this system (*e.g.*, the pelvic nerve, for relief of pain in pelvic structures); (*c*) section of the corresponding dorsal roots (rhizotomy); (*d*) section of the spinothalamic tract (cordotomy).

Abdominal Sensibility

The free nerve endings of somatic nerves branching in the subperitoneal tissue are pain receptors. Pressure on the parietal peritoneum, therefore, awakens pain which is fairly well localized and accompanied by cutaneous hyperalgesia and contraction of the abdominal wall muscles over the stimulated area. The peritoneum lining the diaphragm has the same sensibility as the diaphragmatic pleura, *i.e.*, the peripheral part is innervated by the intercostal nerves and the center by the phrenic nerve. Pain arising in the periphery is referred to the base of the thorax and the upper part of the abdominal wall on the affected side, over an area covered by the dermatomes of the seventh to twelfth thoracic roots. Pain arising in the central area is referred to the shoulder and lower part of the neck (dermatomes corresponding to the third, fourth, and fifth cervical roots).

The omentum is apparently completely insensitive; the peritoneum covering the duodenum and pancreas, on the contrary, has exquisite sensibility. Traction on the mesentery and stimulation of the mesenteric blood vessels or perivascular tissue cause intense pain referred to the epigastrium and the umbilical region.

The digestive tract from the esophagus at the level of the cricoid cartilage down to the anus is relatively insensitive. In conscious patients in the course of surgical operations it has been handled and cut or cauterized, without awakening pain, if care was taken not to pull on the gut. Pain can be aroused from the upper part of the esophagus and the anal canal by application of mechanical stimuli, and cutting the esophagus near the cardia may give rise to a sensation of heartburn. Certain lesions, such as ulceration of the stomach or duodenum and inflammation of the appendix, give rise to intense pain. Sudden distention or intense contraction of the gut at all levels causes pain (colic) which is referred to the epigastrium or the supraumbilical region when it arises in the stomach or small intestine and to the infraumbilical region when it comes from the large intestine. The pain is felt toward the midventral line, even when the affected gut is situated laterally, because the digestive tract develops in the embryo from median and not from lateral structures. Appendicular pain is at first referred to the epigastrium and the umbilical region. Later, when the inflammatory process has invaded the parietal peritoneum, it is localized in the right iliac fossa.

Pain fibers from the stomach form part of the major splanchnic nerves and enter the spinal cord in the dorsal roots from the sixth to the ninth thoracic segments, mainly in the seventh and eighth. Those of the small intestine also form part of the splanchnic nerves but enter the cord

at a lower level (ninth to eleventh thoracic segments). The preaortic plexus and lumbar sympathetic chain include pain fibers from the colon, which enter the cord at the level of the twelfth thoracic and first lumbar segments. Afferent fibers to the rectum are included in the pelvic nerves, and they enter the spinal cord at the level of the second to fourth sacral segments.

Stimulation of the digestive tract by adequate stimuli provokes digestive secretory and motor reflexes, which normally do not enter the sphere of consciousness (see Sec. IV, Digestion).

The parenchyma of the liver, kidney, and spleen can be handled, cut, cauterized, or crushed without evoking any sensation. Parenchymatous lesions of these organs are painless. Sudden distention by vascular congestion, such as occurs in liver stasis of cardiac insufficiency, is painful; the impulses probably arise from the fibrous capsule of the liver, owing to sudden stretching. Gradual distention of the capsule by hepatic hypertrophy, tumors, or hydatid cysts is not painful.

The pedicles of these organs are sensitive to pain. Thus distention or inflammation of the gall bladder and bile ducts is very painful and provokes severe reactions (nausea, vomiting, contraction of the upper right quadrant of the abdominal wall, etc.). The gall bladder, however, can be cut or cauterized without evoking pain.

Pain fibers from the bile ducts and gall bladder are found in the major splanchnic nerves. They enter the spinal cord with the dorsal roots of the sixth to the ninth thoracic nerves.

The renal pelvis and ureter and the neck of the bladder are sensitive to pain, which is provoked when these structures are distended. Renal "colic," such as is caused by renal stone, consists of intense pain in the lumbar region irradiated down the flank to the scrotum. The cremaster is contracted and the testicle is raised. The testicle is hypersensitive to pressure, and there is an area of hyperalgesia extending slightly below the area of referred pain. The pain fibers of the kidney and urethra form part of the major and minor splanchnic nerves; they enter the cord at the level of the twelfth thoracic and the first and second lumbar segments.

Pain fibers for the fundus of the bladder form part of the superior hypogastric plexus and enter the cord in the lumbar segments. Those for the neck of the bladder are innervated by fibers in the pelvic nerve (parasympathetic), which enter the cord at the second to fourth sacral segments. Pain from the bladder is referred to the suprapubic region, the perineum, and the penis.

Sensibility of the Heart and Blood Vessels

In the adventitia and media of the arch of the aorta, the innominate artery, and the carotid sinus, there are differentiated receptors, similar to the Golgi-Mazzoni organs of the tendons, which respond to changes in tension. These receptors are innervated by the aortic nerve (de Cyon and Ludwig's depressor nerve), the fibers of which form part of the vagus in man, dog, and other species, and Hering's sinus caroticus nerve, which is a branch of the glossopharyngeal. The cells of these fibers are in the sensory ganglia of the vagus (g. nodosum) and glossopharyngeal (g. petrosum) respectively, and they end in the sensory nuclei of the vagus and glossopharyngeal (nucleus of the tractus solitarius) in the medulla. When the walls of the aorta and carotid sinus are stretched, as occurs when the arterial blood pressure rises, the tension receptors discharge impulses along the afferent fibers at a rate proportional to the degree of tension, and reflexes are provoked that cause a fall in blood pressure and transitory apnea (see "Vascular reflexes," pages 195 and 296).

Fibers of the aortic and sinus caroticus nerves also end in the aortic and carotid body respectively. These bodies are chemoreceptors sensitive to changes in the chemical composition of the blood, such as variations in the partial pressures of oxygen and CO_2, pH, ionic equilibrium, certain drugs, etc. When these nerve endings are stimulated, cardiovascular and respiratory reflexes are provoked (see pages 195 and 296).

Certain fibers of the vagus nerve end around the roots of the large veins in the right auricle. These fibers are sensitive to changes in venous and auricular pressure. Thus an increase in venous pressure causes an increase in heart rate (Bainbridge's reflex), and a fall in pressure, such as occurs when there is severe hemorrhage, stimulates the vasomotor center, and prevents a fall in arterial blood pressure.

The intima of the arteries is insensitive to most stimuli. Pain evoked by intra-arterial injection of caustic substances does not arise in the arteries but in the periphery. Manipulation of the arteries is painful because of trauma to

the nerves surrounding the artery. In the neighborhood of the mesenteric arteries there are pacinian corpuscles which are stimulated by distention of the arteries, and impulses are thus discharged along afferent fibers included in the splanchnic nerves.[1] The rate of discharge is related to the degree of distention. The response may be an increase or a fall in blood pressure.

Most of the afferent nerves contain fibers that on stimulation evoke pressor reflexes. There are also fibers in the nerves of the blood vessels, skin, and viscera that on stimulation provoke a reflex fall in blood pressure through vasodilatation in the splanchnic area. To observe their effects the pressor fibers must first be depressed by cooling the nerves, or weak stimuli must be applied.

The normal activity of the heart evokes no sensations. Vigorous cardiac activity, such as occurs during physical exercise or emotional stress, or an abnormally fast rate (paroxysmal tachycardia), or a strong contraction, such as the one following an extrasystole, may be perceived and is usually referred to the thoracic wall, the epigastrium, or the lower part of the neck. These sensations do not arise in the heart but in the thoracic wall. The heart itself is insensitive to manipulation and to trauma, and even severe lesions of the endocardium are painless, but ischemia due to spasm or thrombosis of a coronary artery causes the painful syndrome known as angina pectoris (see page 161). It consists of intense pain, accompanied by a sensation of anguish and imminent death. There is deep retrosternal pain, which radiates to the upper part of the left side of the thorax and down the inner aspect of the left arm to the left hand, also up the neck to the angle of the jaw. A sensation of thoracic constriction and marked precordial oppression occurs as a result of contraction of the intercostal and retrosternal muscles (visceromotor reaction). Skin over the painful area is hypersensitive (viscerosensitive reaction), and there are several visceral reactions such as constriction of the blood vessels of the face and neck, sweating, salivary secretion, etc. (viscerovisceral reactions). The painful impulses of angina pectoris are conducted by fibers of the middle and lower cardiac nerves, which pass through the middle cervical, stellate, and upper thoracic sympathetic ganglia to the dorsal roots of the first five or eight thoracic nerves (Fig. 81, page 161). The pain of angina pectoris is temporarily relieved by anesthetizing with novocaine the stellate ganglion and the upper thoracic sympathetic chain down to the fourth or fifth thoracic segment. It is permanently abolished by extirpation of these nerves. The pain is not felt even when there is severe or fatal coronary thrombosis, but the nonpainful reactions, such as dyspnea, vasoconstriction in the skin of the face, palpitations, etc., persist because these reactions are mediated by fibers other than those conducting pain impulses.

On the other hand, stimulation of the stellate ganglion reproduces the pain of angina pectoris.

The mechanism of anginal pain is similar to that of pain provoked by the contraction of skeletal muscle after the circulation has been stopped. Here also there is probably accumulation of metabolites which stimulate the free nerve endings in the myocardium (see "Pain—Mechanism of Excitation," page 887).

The visceral and parietal pericardium are completely insensitive, except for the lower part of the fibrous pericardium, where stimulation provokes pain and cardiovascular reflexes. The pain of pericarditis is due to tension in those cases in which the effusion greatly distends the pericardium (dull precordial pain and oppression) or may be due to disturbances of the coronary circulation (anginal pain) or to extension of inflammatory lesions to the parietal pleura.

Pulmonary and Pleural Sensibility

The parenchyma of the lung and the visceral pleura are insensitive to manipulation and to trauma, but there are stretch receptors that respond to variations in tension and provoke respiratory reflexes (see "The Hering-Breuer reflex," page 292), innervated by the vagus nerve.

The respiratory mucosa (larynx, trachea, and bronchi) is innervated by the vagus. It is sensitive to mechanical stimulation (foreign bodies) and to irritating agents (gases), which may evoke pain and respiratory reflexes such as coughing, apnea or hyperpnea, and laryngeal spasm.

The serous surface of the parietal pleura can be stroked lightly without provoking pain or any other sensation, but the subpleural tissue is innervated by branches of those intercostal

[1] Gammon, G. D., and D. W. Bronk, *Am. J. Physiol.*, **114**, 77, 1935.

nerves which conduct pain impulses. Pressure exerted on the parietal pleura by means of a blunt instrument provokes localized pain, accompanied by hyperalgesia of the skin and deeper structures lying over the stimulated area. There is also contracture of the ipsilateral thoracic muscles, which reduces the excursion of the respiratory movements.

The diaphragmatic pleura is sensitive to mechanical stimulation and inflammatory irritation. Pain evoked from the peripheral part of the diaphragmatic pleura is referred to the base of the thorax on the stimulated side; *i.e.*, to the dermatomes of the sixth to the twelfth thoracic roots. Cutaneous and deep hyperalgesia and muscular contracture are also observed. The central portion of the diaphragmatic pleura is innervated by the phrenic nerve. Pain evoked by stimulation of this area is referred to the base of the neck and shoulder on the affected side, *i.e.*, to the territory of the third, fourth, and fifth cervical dermatomes. Cutaneous and deep hyperalgesia in this region and contracture of the neck muscles are also observed.

HEADACHE[1]

The different types of headache vary in intensity, duration, and the part of the head to which they are referred. Some have a throbbing or pulsating quality, but in all there is the persistent diffuse pain implied in its name and common to all forms of deep pain.

Headache always arises in structures situated in the head or the neck. Pain from other parts of the body is not referred to the head, except in severe and prolonged attacks of angina pectoris, in which case the pain irradiates up the neck to the jaw.

The soft structures covering the face and cranium are sensitive to pain, but the bones of the cranium, including the diploe, the emissary veins, the greater part of the dura, pia, arachnoid, and ependymal membranes, and the cerebral parenchyma are not. The following intracranial structures are sensitive to pain: (*a*) the tissues surrounding the venous sinuses and their tributary veins; (*b*) the arteries in the dura and the cerebral arteries, especially those at the base of the brain; (*c*) the dura of the base; (*d*) the roots of the fifth, ninth, and tenth cranial nerves and the first cervical nerve. Stimulation of endocranial structures evokes only pain sensation, except when an afferent path or center is stimulated, in which case the corresponding specific sensation is evoked.

[1] WOLFF, H. G., *Proc. Assoc. Res. Nerv. Ment. Dis.*, **23**, 173, 1943.

The fifth nerve conducts pain impulses from the face, scalp, and orbit and from the intracranial structures above the cerebellar fossa. Pain arising from endocranial structures in this area is referred to the frontal and parietal regions. The ninth and tenth cranial nerves and the first three cervical nerves conduct pain impulses from the cerebellar fossa and the base of the brain. Pain arising in this area is referred to the occipital region. All these facts have been ascertained in the course of endocranial operations performed under local anesthesia on conscious patients.

Headache of intracranial origin is due to dilatation of the arteries or veins, traction, compression or displacement of sensitive structures, inflammatory processes, and lesions of the sensory nerves (fifth, ninth, tenth, and upper cervical nerves).

Arterial dilatation. Injection of histamine provokes throbbing pain in the head due to dilatation of intracranial arteries, especially those of the base of the brain and the branches of the internal carotid. The intensity of the headache increases with the amplitude of pulsation. If the intracranial pressure is raised by injecting saline into the subarachnoid space, pulsation and headache diminish.

Arterial constriction does not provoke headache. Thus, adrenaline applied to the middle meningeal artery may provoke maximal contraction without causing pain, while dilatation of this artery is painful.

Changes in intracranial pressure. If a needle is introduced into the subarachnoid space and connected to a water manometer, fluctuations in intracranial pressure can be followed and correlated with sensations experienced by the subject.[1] A fall in pressure caused by extraction of a small amount of cerebrospinal fluid (about 20 cc.) provokes headache owing to dilatation of the venous sinuses. The intensity of the headache is increased by exerting pressure on the jugular veins (thus increasing intracranial venous pressure) or by tilting the subject to an upright position (thus increasing the difference between intracranial and venous pressures).

[1] KUNKLE, E. C., B. S. RAY, and H. G. WOLFF, *Bull. New York Acad. Med.*, **18**, 400, 1942.

On the other hand, headache is relieved by subarachnoidal injection of saline. Intracranial pressure changes provoke headache by causing tension to be exerted on perivascular tissues that are sensitive to pain.

Cerebral tumors. Pain in cases of cerebral tumor is due to compression or traction exerted on sensitive structures, either directly by the tumor or by the displacements it causes.

Migraine. This is a periodic form of severe headache limited to part of the head or face. It is usually preceded by scotomata (blind spots) and other visual phenomena in the opposite eye field to the side on which pain will be felt. In certain cases auditory, gustatory, or cutaneous sensations instead of visual ones precede the onset of headache. Photophobia, nausea, and sometimes vomiting form part of the syndrome. Migraine is due to dilatation of the branches of the external carotid artery and occasionally of the internal carotid (anterior meningeal, supraorbital, or superficial frontal arteries). Severity of headache is related to the amplitude of pulsation of these arteries. Vasodilating substances increase the intensity of migraine, and it is relieved by vasoconstricting drugs such as ergotamine. Digital compression of the temporal or other superficial arteries relieves the pain in the territory supplied by the compressed artery, while ligation suppresses it. Scotomata and other sensory signs experienced before the headache begins are due to intense transitory vasoconstriction and anemia of the nerve centers; they disappear on administration of amyl nitrite (a vasodilating drug) if the systemic pressure is not lowered at the same time, in which case they are accentuated. Migraine appears to be an allergic reaction; the attack may be brought on by the release locally of a vasodilating substance.

Headache of ocular origin.[1] Disturbances in the functions of the eye are a frequent cause of headache. The conjunctiva is sensitive to touch, cold, and pain; the cornea is sensitive to pain, and its peripheral parts are also sensitive to cold. Stimulation evokes pain well localized on the site where the stimulus is applied; intense and prolonged pain irradiates to the periorbital area and the territory innervated by the ophthalmic branch of the fifth nerve. Photophobia also occurs. The eye muscles are sensitive to stretching, which causes intense pain referred to the depth of the orbit.

Intraocular hypertension causes localized pain in the affected eyeball, the intensity being in direct relation to the degree of hypertension. It can be intolerably severe, and it irradiates to the whole territory innervated by the ophthalmic branch of the fifth nerve. A decrease in intraocular pressure relieves the pain, which disappears when the pressure falls to a normal level (20 to 30 mm. Hg).

Abnormalities in refraction cause headache owing to sustained contraction of the muscles of accommodation. At first pain is referred to the periorbital area, but later to the back of the head, owing to prolonged reflex contraction of the neck muscles. Hypermetropia and astigmatism are frequently the cause of headache, but myopia does not cause headache. Disturbances in the balanced action of the external musculature of the eye cause headache referred to the occipital region, due to sustained contraction of the neck muscles.

Irritation of the iris by mechanical agents or inflammation causes pain in the eyeball, irradiating to the whole territory of the ophthalmic nerve. In these cases a particular form of photophobia is experienced; there is no hypersensitiveness to light, but changes in illumination cause movements in the iris (see "Pupillary reflexes," Chap. 76) which are extremely painful and induce the patient to avoid light. Drugs such as atropine, which cause cycloplegia (paralysis in the mechanism of accommodation), relieve this form of photophobia, because the iris is immobilized.

True photophobia is the response to an excess of light; the eyes are shut or the palpebral fissure is reduced by contraction of the orbicularis, and the frequency of blinking increases. In normal subjects, a sudden increase in illumination provokes this response, but there is no pain, only a sensation of discomfort. In certain cases the threshold for light is lowered, and photophobia occurs when there is a degree of illumination well tolerated by normal subjects. This condition is due to irritation in the structures innervated by the ophthalmic nerve and, in some cases, to irritation in the territory of the maxillary nerve (nasal disturbances). Irritation of the conjunctiva or cornea is the most frequent cause of photophobia; it disappears if the conjunctiva is rendered insensitive by a local

[1] Eckart, L. B., J. M. McLean, and H. Goodell, *Proc. Assoc. Res. Nerv. Ment. Dis.*, **23**, 209, 1943.

anesthetic. Irritation of periorbital structures or the meninges innervated by the ophthalmic nerve also causes photophobia. In cases of disturbances in the pathway of the pupillary reflex to light (*i.e.*, when there is an Argyll-Robertson pupil) photophobia does not occur. Nerve impulses caused by irritation travel along the trigeminal nerve to its main nucleus, which is closely connected with the colliculi, and also to the nucleus of the spinal tract of this nerve, connected with the visual pathway by the spinotectal tract. These impulses are summated with those from the retina; thus the sensations aroused by light are increased. In summary, photophobia is due to summation of impulses from the retina provoked by light and impulses provoked by irritation in the territory of the fifth nerve. Photophobia ceases when the irritating agent is removed.

Nasal and paranasal headache.[1] Stimuli applied to the nasal mucosa and the structures connected with the nasal cavities (sinuses, etc.) arouse almost exclusively painful sensations. Pressure exerted on the floor of the nasal cavities or on the septum, however, is felt and recognized as such, but accompanied by a sensation of discomfort. The most sensitive areas are the conchae and the openings of the sinuses; the cavities of the latter are much less sensitive. The maxillary nerve innervates the nasal and paranasal cavities; therefore section of this nerve or of the fifth nerve, of which it is a branch, causes anesthesia of the whole area. When there is intense and prolonged pain, it irradiates and is referred to other areas innervated by the maxillary nerve and then to the territory of the ophthalmic nerve. This is usually accompanied by cutaneous hyperesthesia and vasodilatation, lacrimal and salivary secretion, and in some cases, photophobia and mydriasis.

Pain is referred as follows:

1. Pain arising from the inferior concha is referred to the upper dental arch. It irradiates to the zygoma, or to the orbit if its origin is the posterior part of the concha.
2. Pain arising from the middle concha is referred to the zygoma and irradiates to the temporal area.
3. Pain arising from the upper concha is referred to the medial wall of the orbit and irradiates to the neighboring parts of the frontal area and the nose.
4. Pain arising from the opening of the maxillary sinus is referred to the upper dental arch and the infraorbital region, irradiating to the forehead.
5. Pain arising from the nasofrontal duct is referred to the internal margin of the orbit and irradiates to the forehead, to the nose, and along the lower margin of the orbit to the zygoma.
6. Pain arising from the ethmoidal cavity is referred to the inner margin of the orbit and the posterior part of the upper dental arch.
7. Pain arising from the sphenoidal sinus is referred to the vertex of the head.

[1] McAuliffe, G. W., H. Goodell, and H. G. Wolff, *Proc. Assoc. Res. Nerv. Ment. Dis.*, **23**, 185, 1943.

Nasal pain is increased by circumstances that cause passive or active congestion of the nasal mucosa, such as pressure on the jugular veins, a draught of cold air, emotional stress, sexual excitation, menstruation, etc.

Headache arising from disturbances in the ear. The external auditory canal has only slight sensitiveness to pain; the tympanic membrane and the mucosa of the middle ear are highly sensitive; the eustachian tube with its opening in the pharynx is less sensitive. These structures are innervated by the glossopharyngeal, and pain arising in them is felt in the depth of the tissues, with little or no irradiation. If it is intense and prolonged, there is pain in the back of the head and shoulders owing to sustained contraction of the neck muscles, as occurs in all forms of prolonged headache.

Headache of other origin. Headache accompanying fever is due to dilatation of the cerebral arteries. Headache of persistent constipation has been attributed to the absorption of toxic substances from the bowel, but there is no evidence that this is the cause. The fact that this type of headache ceases as soon as the bowels are evacuated suggests it is due to a reflex vasodilatation of the cranial vessels.

PROPRIOCEPTIVE SENSIBILITY

Variations in pressure and tension in the depth of the tissues are of great importance for the regulation of posture and movement and for acquiring knowledge of the position in space, size, shape, weight, and resistance of objects in the environment. The receptors awakening these sensations are placed in Sherrington's proprioceptive field. They are found among the fibers of skeletal muscles; on tendons, fasciae, and periosteum; in subcutaneous, subpleural, and subperitoneal tissue; in the mesentery; and on synovial membranes. There are several types of proprioceptors: (*a*) muscle spindles; (*b*) Golgi

tendon organs; (*c*) pacinian corpuscles; (*d*) the labyrinth or vestibular organ placed within the temporal bone, which is formed by the semicircular canals, and the otolith organs (saccule and utricle). The labyrinth is stimulated by the position and movements of the head. It will be considered in detail in Chap. 81.

FIG. 395. Proprioceptive receptors of striated muscles: *A*, Golgi tendon organs, stimulated by stretching the tendon, actively by muscular contraction or passively by an outside agent; *B*, muscle spindle, stimulated by stretching the muscle; when the muscle contracts it ceases to discharge. *a* and *b*, muscle fibers; *m*, spinal cord; *s*, receptors. (*Fulton, J. F., and J. Pi-Suñer, Am. J. Physiol., vol. 83, p. 556, 1928.*)

Tension receptors. The muscle spindles and Golgi tendon organs are stimulated by tension.

Muscle spindles. (Fig. 395*B*) are made up of several coarsely striated fine muscle fibers (intrafusal muscle fibers) enclosed within a spindle-shaped connective-tissue sheath. They are innervated by several nerve fibers:

1. Several small myelinated fibers which degenerate after the corresponding ventral roots have been cut. They belong to the small-nerve motor system (see "Innervation of muscle fibers," page 790). Several spindles receive branches from the same fiber, so that activity in one fiber influences more than one spindle.
2. A large afferent fiber of the α type which loses its myelin sheath and ends around the middle of the intrafusal fibers, forming ribbonlike "annulospiral" endings.
3. A medium-size myelinated afferent fiber which ends by many fine branches on the intrafusal muscle fibers forming "flower-spray" endings.
4. An unmyelinated fiber of the C type similar to the unmyelinated fibers of the capsulated cutaneous receptors.

Golgi tendon organs (Fig. 395*A*) are situated on the tendons near the insertion of the muscle fibers. They are formed by several fine tendon fibers surrounded by a connective-tissue capsule and are innervated by a large afferent fiber which loses its myelin sheath and branches into many fine terminals among the encapsulated tendon fibrils.

When changes in tension occur in the muscle during contractions or passive stretching, discharges are registered in the afferent nerves. Matthews[1] has shown that in the cat individual stretch receptors fall into two classes:

[1] MATTHEWS, B. H. C., *J. Physiol.*, **78**, 1, 1933.

1. *A* receptors (muscle spindles), which discharge 5 to 10 impulses per second when the muscle is relaxed and discharge at a lower rate or cease to discharge when the muscle contracts (Fig. 396*A*). They have a low threshold to stretch, and when the muscle is passively stretched they discharge at a higher rate; a sudden pull may cause discharges of up to 500 impulses per second. Stimulation of the fine efferent fibers increases the rate of afferent discharge from the muscle spindles. At low muscle tensions this effect of the fine efferent fibers is less than at higher tensions, but it can nevertheless compensate for the effect of decrease in tension which occurs during ordinary muscular contraction. Facilitation of afferent discharges is apparently due to mechanical changes caused by contraction of the intrafusal muscle fibers. After stimulation of the small-nerve fibers or on release from stretch, there is a period of depression lasting several seconds during which afferent discharge may cease or takes place at a lower rate, returning later to the rate corresponding to the degree of tension.[1] Some of the *A* receptors (Matthews's A_2 type) discharge during contraction evoked by excitation of large efferent fibers; these fibers do not innervate the spindles and therefore are not the cause of the discharge, which is due to mechanical (tension) changes in the muscles. Stimulation by stretching causes depolarization of the sensory nerve endings in the muscle spindle and a local potential which spreads electrotonically along the axon. This local potential varies directly with the rate and amplitude of stretching and gives rise to repetitive discharge in the sensory nerve. At low tensions there is a low frequency of discharge, and propagated impulses alternate with monophasic action potentials which do not spread. In the ascending limb of the propagated potentials two or three steps may be seen due to these "prepotentials" of abortive impulses. Release from stretch is associated with a positive potential, which accounts for the poststimulation depression mentioned above.[1]
2. *B* receptors (Golgi tendon organs), which discharge when the muscle is stretched passively, and discharge at a higher rate when the muscle contracts, especially when it

Fig. 396. Left: A-type discharge. Recording from single spindle afferent in dorsal root; initial muscle tension of 10 gm. causes base-line discharge (*BL*). Lower record: maximal stimulus to muscle nerve causes cessation of discharge during contraction. Time base, 50 c.p.s. Right: B-type discharge. Recording from single afferent of another muscle; muscle tension of 20 gm. causes no discharge (*BL*). Maximal stimulus to nerve causes burst of discharges during contraction. (*Kuffler, S. W., and C. C. Hunt, Research Publ., A. Nerv. & Ment. Dis., Vol. 30, 24, 1950*).

[1] Hunt, C. C., and S. W. Kuffler, *J. Physiol.*, **113**, 283 and 298, 1951.

[1] Katz, B., *J. Physiol.*, **111**, 248 and 261, 1950.

shortens (Fig. 396*B*). Their stretch threshold is higher than that of the muscle spindles.

Muscle spindles are placed in parallel with the muscle fibers, and Golgi tendon organs in series. Passive stretch increases the rate of discharge from both types of receptors; contraction increases the discharge of Golgi tendon organs and diminishes that of muscle spindles. Afferent discharges from muscle spindles can evoke discharges from the motor neurons corresponding to the same muscle and can facilitate monosynaptic reflex responses (see "The stretch reflex," Chap. 81) in the same and in synergistic muscles. Activity of the large motor-nerve fibers is usually accompanied by activation of the small-nerve fibers. Afferent discharges from Golgi tendon organs inhibit the motor neurons to the same and synergistic muscles, and facilitate responses in the antagonistic muscles. The over-all effect tends to produce smooth muscle contractions.[1]

In the frog the muscle fibers and muscle spindles are innervated by branches of the same fibers; both systems are "coupled" (see "Innervation of muscle fibers," Chap. 67). In mammals muscle fibers and spindles are innervated by two separate systems of motor fibers, large and small. The pattern of spindle innervation is such that it allows finely graded excitation. This is associated with the importance in mammals of the myotatic reflex which arises in the muscle spindles.

Pressure receptors. The pacinian corpuscles (Fig. 390*f*) discharge when submitted to pressure. Stimulation of a single pacinian corpuscle of the cat's mesentery by mechanical deformation evokes impulses in the nerve with a threshold movement of 0.5 μ in 100 μsec. The duration of the maintenance of the displacement has no influence on the response; change, not duration, of the mechanical deformation is the important factor. Latency of response can be reduced from 1.5 msec. for the threshold stimulus, to 0.5 msec. by increasing the strength of the stimulus. The pacinian corpuscle acts simply as a means of applying mechanical stimulus to the axon.[2] These end organs are found in the deep layers of the skin; there are large numbers in the subcutaneous tissue of the palm and sole. They are also distributed on the fasciae and periosteum, on the tendons near their insertion at the joints, in the subpleural and subperitoneal tissue, in the mesentery close to the arteries, and on the surface of the joints. They are innervated by large fibers.

Pressure receptors are tested by placing weights on the body surface and asking the subject to differentiate the heavier from the lighter. Appreciation of tension can be examined by exerting a pull of known strength on a limb, or by lifting weights. The sense of active movement and position is explored by asking a blindfolded subject to place a limb in a certain position (*e.g.*, the first finger on the tip of the nose) or to describe the position of a limb. The sense of passive movement is examined by moving a joint, *e.g.*, by grasping a finger and flexing or extending it. Goldscheider noted many years ago that the minimum angle and rate of displacement perceived is smaller at the proximal than at the distal joint of a limb.[1]

Proprioceptors are explored in animals by means of the conditioned-reflex technique (see Chap. 86), or by making them learn to distinguish weights or strains in order to obtain a "prize" (food), or by the "placing" and "hopping" reactions. The "placing" reaction is evoked by bringing the paw of a blindfolded animal into contact with the edge of the table on which it is standing; when the leg is released, the animal places it on the table. The "hopping" reaction is evoked by holding the blindfolded animal so that the weight of the body is borne by only one leg; if the animal's body is brought forward, it will "hop," *i.e.*, displace the leg forward so as to keep it below the body in order to bear its weight.

Adaptation. Muscle spindles adapt very slowly and incompletely; they discharge continuously impulses of low frequency. The pacinian corpuscles have a slightly higher degree of adaptation (Fig. 387). Proprioceptors are not easily fatigued.

Proprioceptive reactions. Impulses discharged by proprioceptors initiate and maintain muscle tonus and postural reflexes. They are of

[1] Kuffler, S. W., *Research Publ., A. Nerv. & Ment. Dis.*, **30**, 24, 1952; Hunt, C. C., *J. Physiol.*, **117**, 359, 1952

[2] Gray, J. A. B., and J. L. Malcolm, *Proc. Roy. Soc., London, s.B.*, **137**, 96, 1949.

[1] The functional significance of this resides in the fact that an error in the execution of a movement is multiplied by the length of the lever. Precise movements, therefore, require greater accuracy of appreciation at the shoulder than at the fingers.

fundamental importance for the correct execution of movements, especially of automatic movements such as those of locomotion.

Proprioceptive impulses may reach the level of consciousness, but they are usually unconscious. They have little or no affective tone, unless they are associated by experience with actions of emotional significance.

Vibratory sensibility. Rhythmic deformation of the skin and deep tissues produced by a vibrating body arouses a peculiar sensation related to the amplitude and frequency of the vibrations of the stimulus. This sensation is tested by placing a vibrating tuning fork on the skin, usually over a bone near the body surface (*e.g.*, the distal end of the radius or the tibia), because the proximity of a resistant plane improves the mechanical conditions of vibration and increases the intensity of the sensation. Vibratory sensibility is not, however, peculiar to bone, and it is a misnomer to call it "osseous sensation"; it can be perceived in fleshy parts far from bony structures. Vibratory sensibility can be examined more accurately with special vacuum-tube oscillators (pallesthesiometers).

Vibration and pressure senses are closely associated. Impulses for both sensations are conducted by large α fibers along the proprioceptive pathway. Disturbances in deep somatic sensibility due to spinal lesions are always accompanied by disturbances in vibratory sensations, but chordotomy (section of the spinothalamic tract) does not affect it. Superficial anesthesia of the skin by cocaine infiltration does not alter the threshold of vibratory sensation, but if the deep tissues are also anesthetized over a fairly large area the sensation is lost. After extirpation of the gasserian ganglion (an operation performed in cases of persistent neuralgia of the trigeminal nerve) the skin of the face is no longer sensitive to touch, temperature, or pain, but vibratory sensations are still perceived because the afferent fibers of the facial muscles remain.

The vibration sense, as tested with a special vibrator giving 180 c.p.s., is highest on the finger tips and lowest on the eyelids. Cooling the skin, strenuous exercise, or general fatigue diminish sensitiveness to vibrations.

No special receptors for vibratory sensation have been identified. It is not a separate modality of sensation but a special pattern of pressure sensation caused by repetitive stimulation without the occurrence of fusion. It is therefore similar to flicker in visual sensations.

Stereognosis. The ability to recognize the shape and size of objects (*i.e.*, the geometric tridimensional form of objects) is called stereognosis. This also is not a particular modality of sensation, but a complex pattern integrated by cutaneous (touch) and proprioceptive sensation. It is usually explored by asking a blindfolded subject to describe or discriminate standardized geometric forms or various-sized objects placed in the hand,[1] by means of which an approximately quantitative statement of the stereognostic ability can be made.

Stereognosis is disturbed by lesions in the pathways of cutaneous or proprioceptive sensibility, but to a certain extent the one type of sensation compensates for losses in the other. Lesions in the parietal lobe cause a considerable defect in stereognosis, even though the particular sensations that contribute to it may be only slightly deficient. In these cases the cortical mechanism of association that synthesizes the different sensations into the stereognostic pattern is lost. The condition is known as astereognosis.

BIBLIOGRAPHY

ADRIAN, E. D., "The Basis of Sensation. The Action of the Sense Organs," Christophers, London, 1928.

ADRIAN, E. D., The Mechanism of the Sense Organs, *Physiol. Rev.*, **10**, 336, 1930.

BISHOP, G. H., Neural Mechanism of Cutaneous Sense, *Physiol. Rev.*, **26**, 78, 1946.

CAPPS, J. A., with the collaboration of G. H. COLEMAN, "An Experimental and Clinical Study of Pain in the Pleura, Pericardium and Peritoneum," Macmillan, New York, 1932.

DAVIS, H., and S. H. BARTLEY, Sense Organs, *Ann. Rev. Physiol.*, **5**, 345, 1943.

HARTLINE, H. K., Sense Organs, *Ann. Rev. Physiol.*, **4**, 445, 1942.

LERICHE, R., "La Chirurgie de la douleur," Masson et Cie, Paris, 1937.

LEWIS, T., "Pain," Macmillan, New York, 1942.

MACKENZIE, J., "Symptoms and Their Interpretation," 4th ed., Shaw and Sons, London, 1920.

MORLEY, J., "Abdominal Pain," Wood, Baltimore, 1931.

PIÉRON, H., "The Sensations—Their Functions, Processes, and Mechanisms," Muller, London, 1952.

SANDERS, F. K., The Cutaneous Senses, *Ann. Rev. Physiol.*, **9**, 553, 1947.

WOLFF, H. G., and J. D. HARDY, On the Nature of Pain, *Physiol. Rev.*, **27**, 167, 1947.

ZOTTERMAN, Y., Special Senses: Thermal Receptors, *Ann. Rev. Physiol.*, **15**, 357, 1953.

[1] FOX, J. C., and W. W. KLEMPERER, *Arch. Neurol. & Psychiat.*, **48**, 622, 1942.

CHAPTER 74

Conduction and Integration of Afferent Impulses

IMPULSES DISCHARGED BY receptors are conveyed to the centers by nerve fibers of different types, according to the modality of the sensation. A first degree of integration takes place at the level of the spinal cord or the medullary and mesencephalic nuclei of the cranial nerve; the results are relatively simple reflex responses. The lower afferent centers relay impulses to higher centers, each modality of sensation having a special path. Some of the proprioceptive impulses pass into the cerebellum, and for these there is a cerebellar level of integration. All impulses traveling to the cortex, including those from the cerebellum, are relayed in the thalamus, where a more complex process of integration occurs. Olfactory impulses are the only exception to this rule; they arrive at the cortex without passing through the thalamus. Impulses from the thalamic nuclei are projected to the cortex, where integration is completed, the most complex patterns of behavior are coordinated, and the level of consciousness is attained.

Experimental and pathologic lesions provoke different disturbances in sensation according to the type of fiber affected or the level at which the lesion occurs. Thus section of peripheral nerves or dorsal roots suppresses sensation in certain areas, disturbs it in neighboring parts, and leaves it unaffected in others (topographic peripheral, or dermatomal, anesthesia). Fibers of one type are more easily damaged than others by asphyxia, narcotics, pressure, etc.; therefore the different kinds of sensation are unequally disturbed by these agents (peripheral modality dissociation). In certain cases of lesions or disease of the spinal cord in which some of the afferent paths are destroyed while others remain intact, modality dissociation is also observed. Lesions in the centers disturb integration at the corresponding level, but may leave the simpler integration carried out at lower levels almost, although not quite, normal.

CONDUCTION OF SENSORY IMPULSES IN AFFERENT FIBERS

The diameter of the fiber conditions not only its velocity of conduction and its contribution to the compound action potential of the nerve, but also its threshold and susceptibility to asphyxia, drugs (cocaine), cold, pressure, etc. If one of these agents is applied to a nerve, not all the fibers cease to conduct at the same time; some are blocked before others. The action potential is modified, and certain modalities of sensation are suppressed while others are not disturbed. By correlating data thus obtained, it has been possible to determine the type of fiber that conducts impulses corresponding to each kind of sensation.

Records of the action potentials of nerves thus blocked have given the following results: (*a*) *asphyxia*, provoked by compressing a limb with a pneumatic cuff, suppresses conduction first in δ fibers, then in other A fibers in order of diameter (the α fibers being the most resistant), and finally in C fibers; (*b*) *cocaine* blocks C and α fibers first, then other A fibers in the same order as asphyxia (*i.e.*, δ to α); (*c*) *pressure* blocks α fibers first, then A fibers in decreasing order of thickness, and finally C fibers. The different modalities of sensation are lost progressively in an order that varies with the method of blocking the nerve (Table 95).

Touch fibers. Tactile sensations are suppressed early in asphyxia but late when cocaine is applied to a nerve; moreover, action potentials

Table 95. Order in Which Modalities of Sensation Are Abolished in Different Types of Nerve Block

Order	*Asphyxia*	*Cocaine*	*Cold*	*Pressure*
1	Touch	Second (slow) pain	Cold	Pressure—vibration
2	First (rapid) pain	Cold	Touch	Touch
3	Cold	Warmth	Pain	First (rapid) pain
4	Pressure—vibration	First (rapid) pain	Warmth	Cold—warmth
5	Warmth	Touch		Second (slow) pain
6	Intense cold	Pressure		
7	Second (slow) pain			

typical of δ fibers are recorded in cutaneous nerves following tactile stimulation of the corresponding area of the skin. Touch impulses are apparently conducted by the largest myelinated fibers in cutaneous nerves, *i.e.*, β fibers, as well as by δ fibers. Probably the different types of receptors are innervated by fibers of different diameter.

Fibers for temperature senses. Krause's corpuscles (cold receptors) are innervated by relatively large myelinated fibers, and when a limb is asphyxiated sensitiveness to cold is lost soon after tactile sensibility and the first or "fast" pain. Intense cold and heat still provoke sensations, which disappear only after more prolonged asphyxia. During the whole process the latent period of thermic sensations increases progressively. These observations suggest that thermic sensations are mediated by a wide range of fibers, varying from δ to C.

Fibers for pain. A painful stimulus applied to a limb evokes a double pain sensation; the first is a well-localized prick, the second a more diffuse and prolonged burning ache. The interval between the two sensations increases as the stimulus is applied to more distal parts. Thus it was found to be 0.9 sec. when the stimulus was applied on the thigh, 1.3 sec. on the knee, and 1.9 sec. on the toe. If the limb is subjected to asphyxia, there is a sudden increase in the latent period of the pain sensation from 0.3 to 1.5 sec. at approximately the same time as the δ fibers cease to conduct. These facts suggest that pain impulses are conducted by two types of fibers, δ fibers and C fibers; moreover, when the skin is burned, action potentials typical of C fibers are recorded in the corresponding cutaneous nerve. It would seem that the first ("pricking") pain is conducted by δ fibers, and the second, or slow, pain by C fibers. No definite statement can be made, however, because the reaction time to a stimulus is subject to so many variations dependent on the condition of the subject, the strength of the stimulus, etc. (see "Reaction time," page 873), that it cannot be considered an accurate method for measuring conduction velocity of nerve fibers. Moreover, Bishop[1] has evoked pricking pain by electrical stimulation of the skin with a single shock, while repetition of the shocks on the same spot evoked pain with the aching quality of the second pain. It is therefore possible that the two kinds of pain are dependent on the pattern of stimulation, and not on the type of receptor and fiber that mediate the sensation.

Fibers for proprioceptive sensibility. Nerves passing into muscles have a large number of afferent fibers which innervate the muscle spindles (α and medium-sized myelinated fibers), the Golgi tendon organs (medium-sized myelinated fibers), and pacinian corpuscles (α fibers). There are also fine unmyelinated afferent fibers, which probably mediate pain sensations. The large α fibers convey proprioceptive impulses; they are very susceptible to asphyxia and to pressure, and when they cease to conduct, proprioceptive sensations are lost. The nature of sensations aroused by impulses conducted by medium-sized afferent myelinated fibers that innervate receptors in muscles and tendons has not yet been determined.

Summary. Fibers of a given type may conduct impulses corresponding to several modalities of sensation, although α fibers apparently convey impulses from proprioceptive receptors only. Each modality of sensation sends impulses by several types of fibers, but no modality occupies fibers in all the "fiber spectrum."

The dorsal roots. (Fig. 397.) The dorsal roots are made up of myelinated fibers of all sizes (from 1 to 20 μ in diameter) and unmyelinated fibers. The latter are at least 40 per cent of the total number, and there are more in the thoracic

[1] Bishop, G. H., *J. Neurophysiol.*, **6**, 361, 1943.

and sacral segments, where the visceral afferents enter the cord. The root divides into several filaments, which open out in a fanlike manner before entering the spinal cord at the dorsolateral sulcus. Each filament divides into a medial division, formed by large and medium-sized myelinated fibers, and a lateral division, formed by unmyelinated and fine myelinated fibers. After entering the cord, the fibers divide into an ascending and a descending branch, which give out several collaterals into the gray matter, where they end synaptically on spinal neurons.

All the fibers in the dorsal roots are afferent fibers. The ratio of the total number of fibers (myelinated and unmyelinated) in a root to the total number of cells (large and small) in the spinal ganglion is 1:1. The presence of efferent fibers in the dorsal root has not been demonstrated (see "Antidromic conduction," Chap. 69).

SPINAL ASCENDING PATHS

The fibers in the dorsal roots conduct impulses from all types of receptors; the roots are anatomic units, not functional ones. In the spinal cord there is a functional distribution of the fibers, *i.e.*, each spinal tract conducts impulses corresponding to certain modalities of sensation.

Proprioceptive paths. (Fig. 398.) Proprioceptors are innervated by large α fibers which form part of the medial division of the dorsal-root filaments. Some of these fibers conduct impulses which follow the thalamic route to the cortex; others pass into the cerebellum.

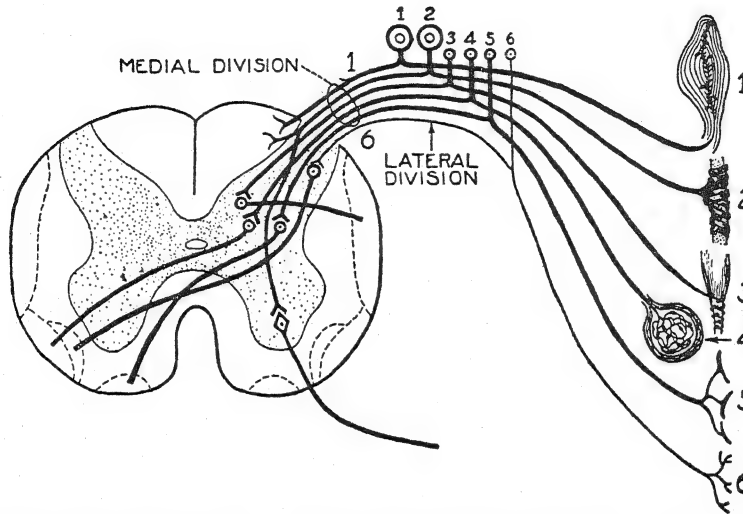



FIG. 397. Origin in receptors and termination in the spinal cord of dorsal-root fibers. 1 and 2, large fibers arising in pacinian corpuscles (1) and muscle spindles (2), and passing to the dorsal column; 3 and 4, fibers arising in tension receptors and ending on dorsal-horn cells, some of the axons of which cross the mid-line and form the ventral spinocerebellar and spinothalamic tracts, while others form the ipsilateral dorsal spinocerebellar tract; 5, fibers arising in tactile receptors and ending on dorsal-horn cells, the axons of which cross and form the ventral spinothalamic tract; 6, fine fibers (pain) ending on cells in the substantia gelatinosa Rolandi; the axons of these cells cross and form part of the spinothalamic tract. (*Fulton, J. F., "Physiology of the Nervous System," 3d ed., Oxford, New York, 1949.*)

Spinothalamocortical propriocebtive path. The fibers in the dorsal root nearest the mid-line form the posterior columns of the spinal cord known as the fasciculus gracilis and the f. cuneatus. These columns are laminated, *i.e.*, the fibers corresponding to each root are distributed in a thin layer or wedge, disposed in a dorsoventral direction. Fibers from the sacral roots are closest to the mid-line, and as the more cephalic roots enter the cord they are placed laterally to those corresponding to more caudal dermatomes (Fig. 400). The fibers send collaterals into the gray substance of the cord, which end on spinal neurons. Some of them end directly on motor neurons of the ventral horn, forming two-neuron monosynaptic arches, and others end on internuncial neurons. The dorsal tracts ascend to the medulla; the fibers near the mid-line end

in the nucleus gracilis, and the more lateral ones in the n. cuneatus. Dorsal-column fibers are relatively large; 16.7 per cent and 10.5 per cent of those in the f. gracilis and cuneatus, respectively, have diameters larger than 6 μ, and the largest are 15 μ in diameter. The speed of conduction is correspondingly great; in the fastest fibers it is 70 m. per sec. There is, however, an abrupt fall in the speed of conduction of impulses from the dorsal roots conveyed in the dorsal funiculi.[1] Thus, afferent impulses from the knee joint are transmitted at 90 to 100 m. per sec. in peripheral nerves, at 40 to 60 m. per sec. in the dorsal tract of the thoracic spinal cord, and at 20 to 30 m. per sec. at the level of the first cervical segment.[1] This may be due in part to the fact that the largest fibers enter almost immediately into the Clarke-Stilling column, and perhaps to a decrease in the size of the fibers as they ascend and give off collaterals.

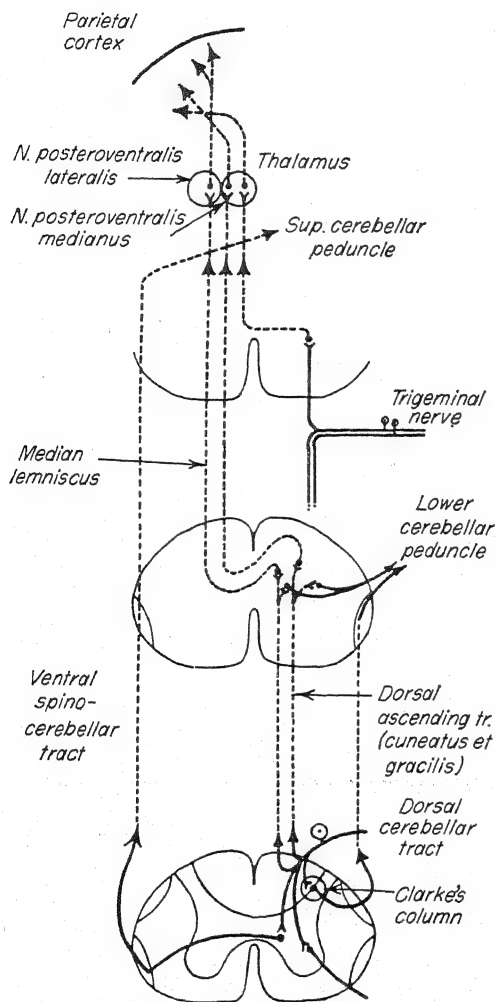



Fig. 398. Proprioceptive pathways.

Second-order fibers, emerging from the gracile and cuneate nuclei, cross to the opposite side and form the medial lemniscus, an ascending tract placed close to the mid-line in the pons and more laterally in the cerebral peduncles. Lamination is continued in the medial lemniscus; fibers from the cuneate nucleus, *i.e.*, those corresponding to the cervical roots, are placed dorsally and those from the sacral roots ventrally. In the cerebral peduncle, the sacral fibers are the most lateral and the cervical fibers are closest to the mid-line.

The medial lemniscus ends in the thalamus, in the lateral part of the posteroventral nucleus. Fibers conveying impulses from the caudal segments end on cells situated laterally to the cells on which the more cephalic segments are projected. The main sensory (mesencephalic) nucleus of the trigeminal nerve sends fibers through the dorsal secondary tract to the medial posteroventral nucleus of the thalamus. Third-order fibers pass from these thalamic nuclei to the parietal cortex. Impulses from the more caudal segments are projected on the upper part, and those from the more cephalic on the lower part of the postcentral gyrus. Cortical representation is, therefore, contralateral and inverted; a similar topographic representation is found for cutaneous receptors.

Impulses conducted along this path give rise to well-localized, discriminative, and conscious sensations. Knowledge is thus acquired of the position of the limbs, active and passive movement, tension, and deep pressure (proprioceptive or kinesthetic sensations). Touch and pressure impulses and those serving to construct complex sensation patterns, such as vibration and stereognosis, are also conveyed along this path.

Spinocerebellar paths. The largest of the α fibers in the dorsal root enter almost immediately into

[1] Gasser, H. S., and T. H. Graham, *Am. J. Physiol.*, **103**, 303, 1933; Grundfest, H., and B. Campbell, *J. Neurophysiol.*, **6**, 275, 1942; Rexed, B., and G. Ström, *Acta Physiol. Scandinav.*, **25**, 219, 1952; Tasaki, I., *Japan. J. Physiol.*, **3**, 73, 1952; Holmgren, B., *J. Physiol.*, **123**, 324, 1954.

[1] Gardner, E., F. Latimer, and D. Stillwell, *Am. J. Physiol.*, **159**, 195, 1949.

the Clarke-Stilling column.[1] The axons of the cells in this column form the ipsilateral dorsal cerebellar tract (Flechsig's tract), which ascends along the dorsolateral margin of the lateral column, enters the cerebellum by the inferior cerebellar peduncle (restiform body), and ends in the cortex of the anterior lobe of the cerebellum (lingula, centralis, and culmen). The fibers in this tract are even larger than those of the dorsal column; almost half have diameters larger than 6 μ. The impulses are also conducted at a higher speed than in the dorsal columns.

Other fibers in the dorsal roots end on cells in the dorsal horn. The majority of the axons of these cells decussate and form the contralateral ventral cerebellar tract (Gower's tract) situated in the ventrolateral margin of the anterolateral column; a smaller number pass into the ipsilateral ventral cerebellar tract. All these fibers ascend into the brain stem and pass into the superior cerebellar peduncles, where they cross over to the opposite side and end in the cortex of the anterior lobe. The projection in the cerebellum is therefore ipsilateral, not contralateral as in the cerebral cortex.

Paths of cutaneous and visceral sensibility. *Spinothalamic tract.* (Fig. 399.) The fine myelinated and unmyelinated fibers in the lateral division of the dorsal-root filaments conduct impulses from touch and temperature receptors and from pain endings in the skin and viscera. These fibers form the tract of Lissauer at the tip of the posterior horn; they ascend for one to three segments and then penetrate into the substantia gelatinosa Rolandi, formed by small cells in the most dorsal part of the posterior horn. The axons of the neurons that conduct pain impulses cross the mid-line in the segment of entry or that immediately above; those conducting impulses from temperature receptors decussate in the three to five segments above the segment of entry; and those conveying tactile impulses cross in many, perhaps all, of the segments up to the medulla. A few fibers apparently do not decussate but pass into the ipsilateral spinothalamic tract; most of these appear to be visceral afferents. Temperature fibers are placed in the anterolateral column in a dorsal position with respect to the pain fibers; touch fibers are placed more ventrally. As in the dorsal column, there is dermatomal distribution in laminae in the spinothalamic tract; fibers corresponding to the cervical dermatomes are placed near the mid-line, next to these the thoracic, and more laterally and dorsally the lumbar and sacral fibers (Fig. 400).

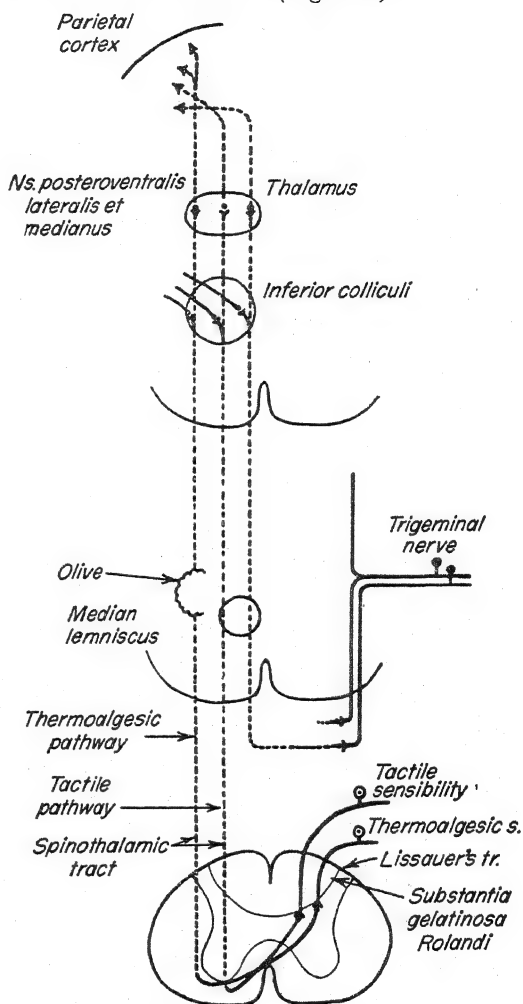



Fig. 399. Spinothalamic tract.

Pain and temperature fibers ascend in the lateral part of the reticular substance, dorsal to the inferior olivary nucleus in the medulla. They are placed laterally and superficially with respect to the median lemniscus in the pons and even more so in the cerebral peduncle. In this part of the tract lamination is also observed; the fibers from caudal segments are placed dorsally with respect to those from more cephalic segments.[1]

[1] Lloyd, D. P. C., and A. K. McIntyre, *J. Neurophysiol.*, **13**, 39, 1950.

[1] Walker, A. E., *Assoc. Res. Nerv. Ment. Dis.*, **23**, 63, 1943.

Some of the fibers in this tract end in the inferior colliculus (spinotectal tract), but most of the fibers pass through these nuclei and end in the posterior ventral nuclei of the thalamus. Fibers from the more caudal dermatomes end in the lateral part of the n. ventralis posterolateralis, and those from more cephalic dermatomes on cells nearer the mid-line. Fibers from the spinal nucleus of the trigeminal nerve (ventral secondary tract) end in the n. ventralis posteromedialis (arcuate nucleus).

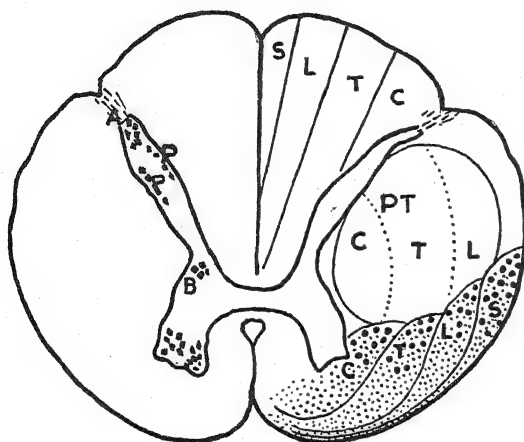



FIG. 400. Diagrammatic section of spinal cord showing systematization of ascending fibers. *S*, fibers from sacral dorsal roots; *L*, fibers from lumbar roots; *T*, fibers from thoracic roots; *C*, fibers from cervical roots; *PT*, pyramidal tract. Large dots represent fibers conducting impulses from temperature receptors, medium-sized dots represent pain fibers, and fine dots tactile fibers. (*Walker.*)

Afferent impulses from the head. Afferent fibers in cranial nerves conveying cutaneous, visceral, and proprioceptive sensations are found mostly in the trigeminal, facial, glossopharyngeal, and vagus nerves. Impulses from proprioceptors in the eye muscles are conducted by fibers of unknown origin.

The trigeminal nerve has large and fine fibers. Most of the former ascend into the main sensory or mesencephalic nucleus and most of the latter descend into the spinal nucleus. This nucleus is a continuation of the substantia gelatinosa Rolandi of the spinal cord, and the trigeminal fibers ending in it convey mainly pain and temperature impulses, and secondarily those of touch. Touch, pressure, and proprioceptive sensibility are served by the fibers passing into the main nucleus. Destruction of the spinal nucleus completely suppresses temperature and pain sensibility in the territory innervated by the trigeminal nerve. The lower part of the nucleus receives fibers from the ophthalmic branch, and the upper part from the mandibular branch. Second-order fibers from the trigeminal nuclei end in the arcuate nucleus (n. ventralis posteromedialis) of the thalamus.

The thalamic nuclei are projected on the postcentral gyrus in the same way as the proprioceptive fibers, *i.e.*, the head and face in the lower part of the gyrus and the more caudal segments in the higher. The perineum and genital organs are represented in the medial aspect of the hemisphere above the falx, *i.e.*, in the paracentral lobule.

Cordotomy. Section of the anterolateral spinal column, *i.e.*, the spinothalamic tract (cordotomy), has been performed in many cases in order to relieve intractable pain. The operation must be bilateral to abolish visceral pain; therefore visceral pain fibers decussate only in part. Cordotomy also suppresses temperature sensations, but it has little or no effect on tactile sensibility. According to Foerster, bilateral cordotomy reduces the number of touch spots and raises their threshold, but the deficiency is detected only by the use of precise methods of exploration. Sensations from the bladder also persist, but those of sexual orgasm are lost. The persistence of bladder sensibility may be due to the dorsal position of the fibers that conduct these impulses, thus causing them to escape section; it is also possible that bladder sensibility is conducted by fibers in the dorsal columns.

Spinal sensory syndromes. Modality dissociation of sensation occurs in cases of spinal disease in which certain tracts are damaged while others remain intact.

Degenerative lesions of the posterior columns (fasciculus gracilis and f. cuneatus) cause a syndrome called locomotor ataxia (tabes dorsalis). Muscle tonus is lost or considerably diminished, owing to the loss of the myotatic reflex (see Chap. 81). Tendon reflexes, such as the knee jerk, are also abolished. The sensation of limb posture is lost, and the patient cannot coordinate voluntary or automatic movements. Visual sensations compensate in part for the deficiency; therefore if the patient closes his eyes he can no longer stand upright, but loses his balance (Romberg's sign). Tactile sensations are considerably

disturbed, especially perception of complex sensation patterns, such as stereognosis.

Destructive lesions in the gray substance of the spinal cord produce the syndrome known as syringomyelia (Greek σύριγξ, pipe, and μυελός, marrow). Cells and fibers of the spinothalamic tract are destroyed and pain and temperature sensations are lost, but tactile and proprioceptive sensibilities remain.

THE THALAMUS

The thalamus is a nuclear mass where all afferent fibers to the cortex, with the sole exception of the olfactory fibers, are relayed. Functional and anatomic connections of the thalamus with the cortex and other centers have been studied mainly by the following methods: (*a*) localized destruction of the thalamic nuclei or the fibers emerging from these nuclei, and examination of serial sections stained by the Marchi method, in order to visualize and follow the path of degenerating fibers; (*b*) localized destruction of cortical areas and examination of retrograde degenerative processes, such as chromatolysis (Nissl's method), which are particularly severe and end in complete atrophy of the thalamic nuclei; (*c*) local application of strychnine (2 per cent solution, mixed with a stain in order to localize the area into which the drug has diffused) and registration of electrical activity in the thalamus on stimulating its afferent paths, or in centers where the thalamic fibers end.[1]

The thalamic nuclei. The thalamus is divided into three parts by the internal medullary lamina, a vertical sheath of white matter which bifurcates at its anterior end.

The anterior nuclei are enclosed between the branches of this bifurcation. The mammillothalamic tract ends in these nuclei, which project to the gyrus cinguli.

The median nuclei are situated between the lamina and the mid-line. The main nuclei are (*a*) the dorsomedian (n. medialis dorsalis), which projects to the prefrontal lobe and the hypothalamus; (*b*) the centrum medianum, (*c*) clusters of cells in the lamina (intralaminar nuclei); (*d*) clusters of cells in the wall of the third ventricle (paleothalamus) connected with the hypothalamus, but with no cortical connections (nuclei of the mid-line).

[1] Dusser de Barenne, J. G., and O. Sager, *Ztschr. f. d. ges. Neurol. & Psychiat.*, **133**, 231, 1911; *Arch. Neurol. & Psychiat.*, **38**, 913, 1937.

The lateral nuclei are placed externally or laterally to the lamina. The main nuclei are (*a*) n. ventralis anterior; (*b*) n. ventralis lateralis, which receives fibers from the cerebellum through the superior cerebellar peduncle and projects to the precentral gyrus (areas 4 and 6); (*c*) n. ventralis posterior, divided into lateral and median (n. arcuatus) nuclei; the former receives the fibers of the median lemniscus and the spinothalamic tract conveying impulses from all the body; the n. arcuatus is a relay station for impulses coming from the head and neck through the trigeminal nerve and the second-order neurons of the trigeminal path; these nuclei project to the postcentral gyrus (areas 3-1-2); (*d*) n. lateralis dorsalis et posterior, the afferent fibers of which are unknown, which projects to the posterior parietal lobule (areas 5 and 7).

The posterior nuclei are situated laterally to the lamina, posterior to the ventral and lateral nuclei. The following are the main nuclei in this group: (*a*) the pulvinar, which projects to the posterior part of the parietal lobule and the peristriate cortex (areas 18 and 22); (*b*) the lateral geniculate body, a relay nucleus in the optic path, which projects to the striate cortex in the occipital lobe; (*c*) the median geniculate body, placed in the auditory path, which projects to the auditory area of the temporal lobe.

The thalamic nuclei can be classified into

1. Cortical relay nuclei placed in the ascending sensory system: the ventralis posteromedialis, which receives the trigeminal lemniscus and projects to the face and neck areas of the somatosensory cortex; the ventralis posterolateralis, which receives the spinothalamic tract (medial lemniscus) and projects to the somatosensory cortical areas corresponding to the leg, trunk, and arm; the lateral geniculate body, on which the optic tract ends and which projects to the striate cortex; the medial geniculate body, receiving the lateral lemniscus and projecting to Heschl's gyrus (auditory cortex); the lateral ventral nucleus, receiving fibers from the brachium conjunctivum and projecting to the precentral gyrus (areas 4 and 6).
2. The diffuse thalamic projection system[1] made up in the cat by five nuclear groups: (*a*) centromedian, (*b*) intralaminar nuclei, (*c*)

[1] Harzl, T. E., and H. W. Magoun, *J. Neurophysiol.*, **14**, 133, 1951.

anterior nuclei, (*d*) anterior ventral nucleus, and (*e*) anterior pole of the reticular nucleus. These nuclei form a functionally interconnected unit, and stimulation of any one of them evokes recruiting waves in the others. Their cortical projections are distinct from those of the previous group; they end in the association areas separating the somatosensory, visual, and auditory cortical areas, *i.e.*, in the prefrontal cortex (overlapping the motor cortex), the cingulate, orbital, and suprasylvian areas. The intralaminar and mid-line nuclei also project to the caudate nuclei[1] which take part in the recruiting responses evoked from the thalamus.

FIG. 401. Diagram of thalamocortical projections. 1, lateral geniculate body; 2, pulvinar; 3, n. lateralis posterior; 4, n. ventralis posterolateralis; 5, n. ventralis posteromedialis; 6, median geniculate body; 7, n. ventralis lateralis; 8, n. medialis dorsalis. (*Walker, A. E., J. Nerv. & Ment. Dis., vol. 85, p. 254, 1937.*)

3. Nuclei with no known cortical projections, *e.g.*, the mid-line nuclei, which do not show lesions of retrograde degeneration after removal of the cortex.

Lamination occurs in the cortical relay nuclei, and it is a prominent feature in the lateral geniculate body. Fibers from the macula of the retina end in the dorsal part of this nucleus in such a way that the corresponding points in the maculae in both eyes project to the same point in the geniculate body. Fibers from the peripheral retina end in the anterior third of the geniculate body; here also there is some degree of topographic organization, but with considerable overlapping of the different areas. Projection to the striate cortex also has a certain degree of point-to-point systematization. The retina is therefore projected first to corresponding points in the lateral geniculate bodies, and then to the occipital cortex.

Systematization of the posteroventral nuclei is such that the fibers of the median lemniscus and the spinothalamic tract end on the lateral nucleus, the different dermatomes placed in order so that the caudal segments are projected laterally to the cephalic segments. Impulses from the face and head are projected to the n. arcuatus (n. ventralis posteromedialis). Topographic organization in the cortex will be considered in detail farther on. The existence of lamination and point-to-point projection does not mean that there is no convergence and divergence; on the contrary, even in the optic pathway of macular vision there is a certain degree of overlapping of the different areas.

The areas of projection of the thalamic nuclei are represented in Fig. 401, and the density of projection in Fig. 402.

The cortex also projects to the thalamic nuclei. Thus the occipital (visual) cortex projects to the lateral geniculate body, the temporal (auditory) cortex to the median geniculate body, and the motor cortex to the lateroventral nucleus. The general rule is that each part of the cortex projects to the nuclei from which it receives impulses. Thus there is the possibility of the establishment of thalamocorticothalamic circuits similar to those which have been found to exist between the cortex and the striatus (see Chap. 82).

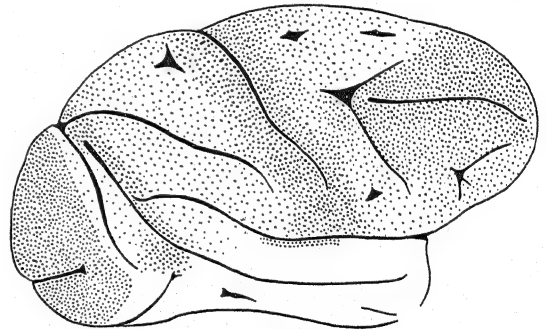



FIG. 402. Relative density of thalamocortical projection in different regions. Diagram of cerebral cortex of macaque. (*Walker, A. E., "The Primate Thalamus," University of Chicago Press, Chicago, 1938.*)

Sensory functions of the thalamus. Thalamic functions have been studied in animals after the

[1] DROOGLEEVER-FORTUYN, J., and R. STEFENS, *Electroencephalog. & Clin. Neurophysiol.*, **3**, 393, 1951.

whole cortex of one hemisphere has been removed. This procedure not only suppresses cortical activity, but also provokes degeneration and atrophy of the thalamic nuclei connected with the cortex, thus causing serious disturbance in thalamic function. Immediately after the operation there is complete anesthesia on the opposite side of the body. Gradually pain sensibility and, to a certain extent, tactile sensibility, are recovered, but accurate localization and discrimination are permanently lost. Visceral sensibility is not greatly altered, while proprioceptive sensations are definitely abolished. In primates the results are more severe than in lower animals. In cases of hemidecortication in man the syndrome is similar to that observed in experimental animals, *i.e.*, complete contralateral anesthesia, with gradual incomplete recovery of cutaneous sensibility.

Partial recovery of superficial sensibility may be due in part to bilateral cortical representation, owing to incomplete decussation of the spinothalamic tract. Thus, for the face there is considerable bilateral representation. That this is not the only reason for partial recovery of sensibility after hemidecortication is proved by the following experiment: the whole cortex is removed from one hemisphere, and after the animal has recovered, the cortex of the other hemisphere is removed. Complete contralateral anesthesia is observed after the second operation, but the partially recovered cutaneous sensibility on the other side is not disturbed.[1] Rudimentary cutaneous sensations (Head's "protopathic" sensibility) can therefore be integrated in the thalamus.

Localized strychninization of the posteroventral nucleus on one side causes bilateral hyperesthesia, more marked on the opposite side than on that of the treated nucleus. This experiment confirms the existence of bilateral representation of cutaneous sensibility. If strychninization is limited to a small part of the nucleus, hyperesthesia is observed only over a certain area of the skin, face, trunk, or limb, according to the part of the nucleus affected by the drug. This experiment confirms the existence of topographic organization revealed by anatomical studies.

Little is known about thalamic integration of visceral sensibility. The connections established by the anterior thalamic nuclei and the posterior hypothalamus (through the mammillothalamic tract) and the gyrus cinguli suggest the existence of a thalamic level of visceral integration, connected in some way with olfactory sensations.

Thalamic projection to cortical association areas (posterior parietal, prefrontal area) suggests that the respective thalamic nuclei play a part in the integration of complex visual and auditory sensations and in stereognosis, which is dependent on the correct integration of cutaneous and deep sensibility.

Thalamic syndromes. Obstruction of the thalamogeniculate artery provokes a thalamic syndrome first described by Déjerine in 1906. The posterior parts of the ventral and lateral nuclei are destroyed, and the following effects are observed: "hemianesthesia, more or less severe for cutaneous sensibility (touch, pain, temperature), but always marked for deep sensibility, with an exaggerated response to stimulation with painful or thermal stimuli, out of proportion to the strength of the stimulus. . . . Pain is felt, frequently intense, on the side where there is anesthesia; it is persistent, has the characteristics of central pain; it is felt deeply, there are paroxysms of lancing pain, which does not respond to treatment with analgesics."[1] The sensations are referred to the side opposite to the lesion, but they are imprecisely localized. There are also motor disturbances in the contralateral limbs (ataxia, choreoatetosis, etc.; see Chap. 82). The face is usually free from disturbances, because the median posteroventral nucleus is seldom damaged. In some cases the lesions are restricted to the entrance of the spinothalamic tract, and cutaneous sensibility alone is disturbed, because the impulses from deeply placed receptors (proprioception, deep pressure, touch) carried by the median lemniscus arrive normally to the thalamus.

Thrombosis of the perforating artery, which supplies the nuclei in the anterior part of the thalamus connected with the striatum, is followed by few sensory disorders, while motor disturbances such as tremor, ataxia, and choreoatetosis are prominent (see "The extrapyramidal motor system," Chap. 82).

Thalamic functions. The thalamus has several functions:

Sensory functions. It is a relay station in the ascending sensory paths. Visceral and somatic sensory impulses are relayed through the posteroventral nuclei; visual impulses through the

[1] WALKER, A. E., and J. F. FULTON, *J. Nerv. & Ment. Dis.*, 87, 677, 1938.

[1] DÉJERINE, J., "Séméiologie des affections du système nerveux," Masson et Cie, Paris, 1914, p. 922.

lateral geniculate body; auditory impulses through the median geniculate body; taste impulses through the median posteroventral nucleus; cerebellar impulses through the lateroventral nucleus; and possibly olfactory impulses through the mammillothalamic tract and the anterior nuclei to the gyrus cinguli.

Corticodiencephalic association, necessary for the integration of complex sensations, visual (pulvinar and peristriate cortex, area 18), auditory (pulvinar and temporal cortex, area 22), and general somatic, including stereognosis (dorsomedian nucleus with the prefrontal cortex, areas 9-10-12, and lateral dorsal nucleus with posterior parietal cortex, areas 5 and 7).

Regulation of cortical activity. The diffuse thalamic projection system, together with the reticular formation of the brain system, plays an important part in regulating the general level of cortical activity. Stimulation of these nuclei may depress the cortex, inhibiting spontaneous movements and evoking the electroencephalographic response of somnolence and sleep, or on the contrary may provoke a condition of alertness (see Chaps. 87 and 88).

CYTOARCHITECTURE OF THE CEREBRAL CORTEX[1]

The cerebral cortex is formed by a series of layers of cells and fibers organized according to a fundamental pattern common to the whole cortex, but with considerable variations in the shape, size, and number of the cells and the distribution of the fibers in the different areas. In all species that have so far been studied, in spite of variations in cell number, form, and size, "what remains constant is the arrangement of the plexuses of dendritic and axonal branches, *i.e.*, of the synaptic articulations through which nerve impulses are transmitted."[2] The comparative study of the cortex in different species is thus made possible.

General structural plan of the cortex. A vertical segment of the cortex may be considered as a "cortical unit" made up of neuron chains "in no way different from the chains of internuncial neurons in any part of the central nervous system."[1] These units, however, are not separated from each other, or placed simply side by side; they overlap and are closely interconnected, so that a single neuron forms part of more than one neuron chain.

Six layers of cells have been found in all parts of the cortex (Table 94 and Fig. 403). The

Table 96. Stratification of the Cerebral Cortex

Parietotemporooccipital Isocortex	*Rhinencephalon (Allocortex)**
I (1). Plexiform layer	I. Plexiform layer
II (2). Small pyramids	II. Star cells
III (3). Medium-sized pyramids	III. Superficial pyramids
IV*a* (4). Star pyramids	IV. Deep pyramids
IV*b* (5). Star cells	V. Small pyramids
V (6). Large deep pyramids	VI. Polymorphic cells
VI (7). Spindles	

* From LORENTE DE NÓ, R., *J. Psychol. Neurol.*, **45,** 318; **46,** 113, 1934.

separation between layers IV*b* (5) and V (6) is clearly marked by a white line formed of myelinated fibers running parallel to the outer surface of the cortex, which is known as Baillarger's external band. It is highly developed in the occipital cortex, where it was first seen by Gennari (Gennari's line) and is the origin of the term "area striata" given to the visual cortex. Baillarger's band divides the cortex into an external lamina with four layers of cells and an internal lamina with two layers. The areas of greatest architectural differences are the relatively simple olfactory cortex and the far more complex pallium. The former, also called "allocortex," has a white external lamina, while the latter, known as "isocortex," has a gray external lamina.

Cortical neurons have been classified into four types according to the distribution of their axons: (*a*) cells with descending axons, some of which penetrate the white substance, which are fibers of projection to subcortical centers or of association with other parts of the cortex; (*b*) cells with short axons, which ramify in the proximity of the cell body (Golgi II cells); (*c*) cells with ascending axons, which ramify in one or more cortical layers; (*d*) cells with horizontal axons (Cajal's cells), which are found in the plexiform layer.

Cortical afferent fibers come from (*a*) the thalamus (thalamocortical projection fibers); (*b*)

[1] A detailed description of the cytoarchitecture of the cerebral cortex will be found in Chap. 15 of Fulton's "Physiology of the Nervous System," 2d ed., written in great part by R. Lorente de Nó.

[2] FULTON, J. F., "Physiology of the Nervous System," 2d ed., Oxford, New York, 1943, p. 280.

[1] *Ibid.*, p. 293.

other cortical areas (cortical association fibers). Projection fibers ramify mainly in layer IV, but all afferent fibers send branches to cells in all the cortical layers; therefore no single layer can be considered as the "receptor layer."

number of terminals from many other neurons, and its axon endings connect with several neurons. Even the axons of projection fibers may send out branches before entering the white matter.

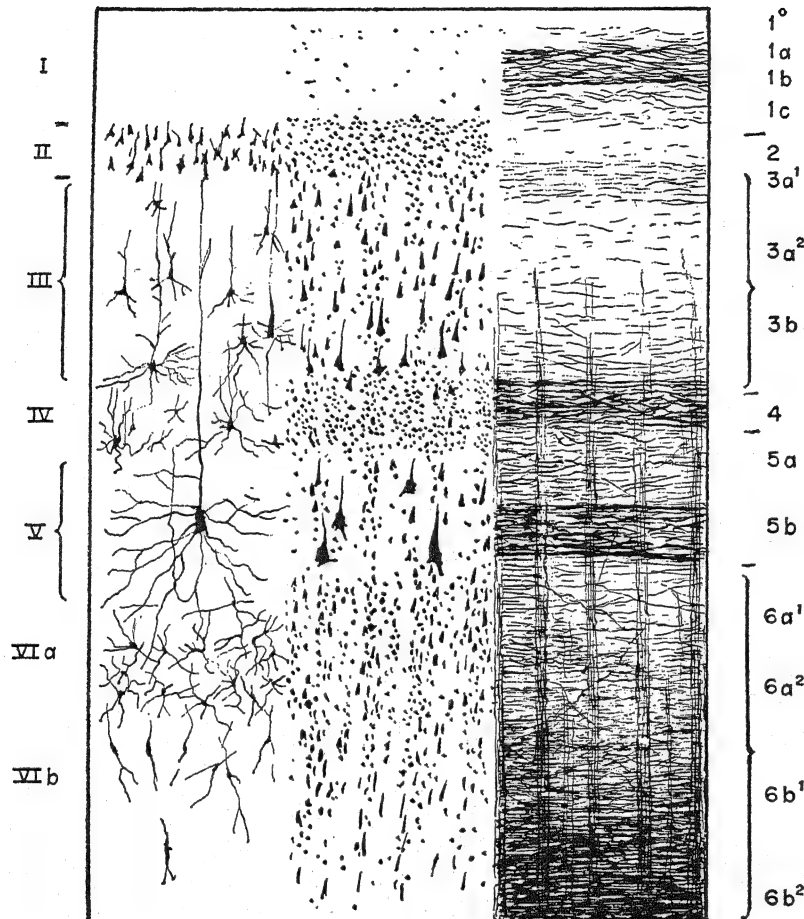



Fig. 403. Cytoarchitecture of the cerebral cortex. On the left, cell layers; on the right, fiber layers. (*After Brodmann.*)

Cortical efferent fibers are sent out from cells in all the layers. The axons of the large pyramids (Betz cells) in the motor cortex (precentral gyrus) form the pyramidal tract. The main efferent fibers from the olfactory and occipital areas arise from the outer layers. Axons of cells in the outer layers ramify mainly within the cortex, but some penetrate into the white matter. No layer can therefore be considered as the "effector layer."

Synaptic connections of cortical cells do not differ from those of other cells in subcortical and spinal centers. Each neuron receives a large

The cell chains may be simple, formed by two cells, with only one synapse in the cortex. Most of the chains are more complex, with many intracortical synapses, and sometimes the chain forms a reverberating circuit. In man the absolute and relative number of neurons with short axons is much larger than in other species; according to Cajal, this is a sign of greater fineness in function.

Cytoarchitectural areas. (Fig. 404.) The differences in cytoarchitecture between two areas are not limited to a single layer, but usually extend to several. The passage from one area to

another is clearly marked; there are no transitional areas. The cytoarchitectural maps of the cortex that have been published are far from being complete, and intensive study of the different parts of the cortex continually shows new subdivisions in areas considered homogeneous. anatomic, but also physiologic and pathologic, significance; hence its importance.

Only the main cytoarchitectural areas will be mentioned here. When considering cortical integration of motor and sensory functions, the respective cortical centers will again be examined.

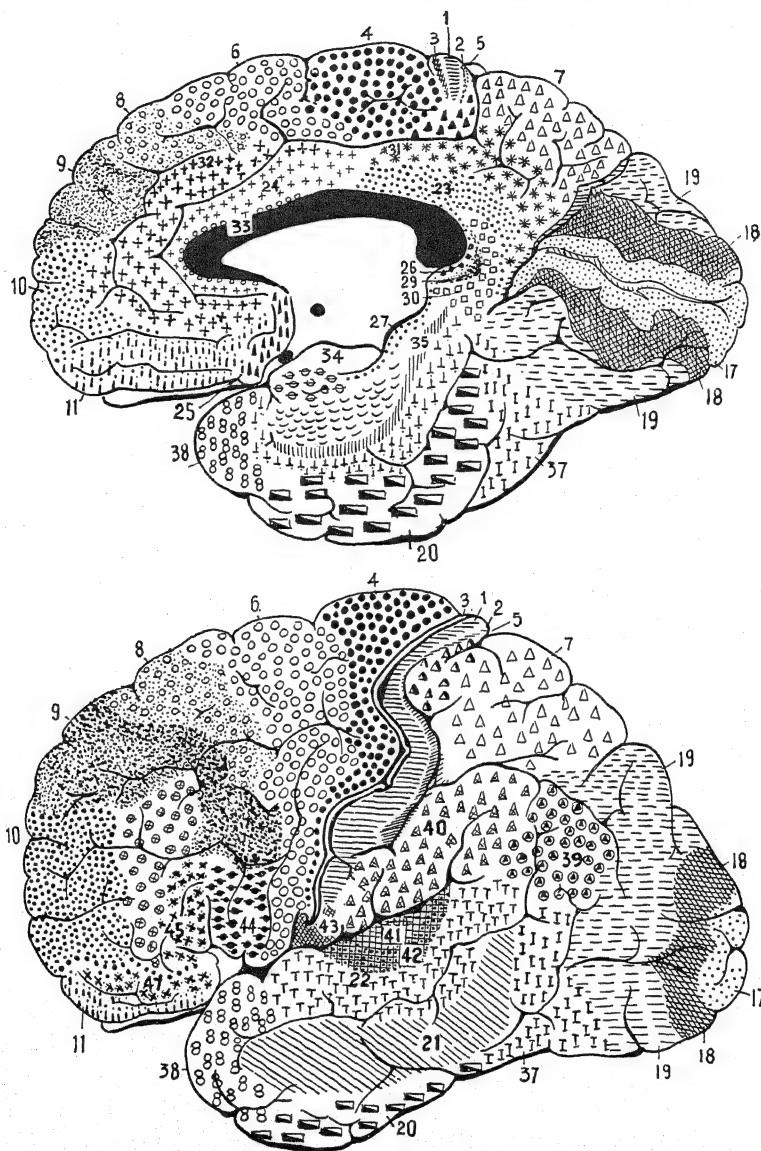



FIG. 404. Brodmann's areas. Diagrams of median (above) and lateral (below) aspects of the human cerebral cortex.

The first maps were constructed by Campbell, Brodmann, and the Vogts; since then there has been considerable and continuous progress in this field of neurology. The division of the cortex into areas with different structure has not only

Frontal lobe. The principal areas are (*a*) the motor area, Brodmann's area 4,[1] situated in the precentral gyrus, whose projection fibers form the pyramidal tract; (*b*) the premotor area (6),

[1] Brodmann's numeration is in general use.

placed in front of the motor area, which forms part of the extrapyramidal motor system; (*c*) the frontal eye field (8), in the posterior part of the second frontal convolution; (*d*) the prefrontal area (9-10-11-12), an association area occupying the rostral part of the frontal lobe.

Parietal lobe. The areas in this lobe are mainly related to general somatic sensations: (*a*) the postcentral area (3-1-2) receives projections from the posteroventral nuclei of the thalamus; (*b*) the preparietal (5*a*) area situated behind the postcentral area is also a sensory area, but contributes to the corticopontine and corticospinal projections; (*c*) the inferior parietal area (7), which includes the supramarginal gyrus (40) and the angular gyrus (39), is an association area in which are integrated somatic auditory and visual impulses of importance in the mechanism of speech.

Temporal lobe. The lower lip of the sylvian fissure (Heschl's transverse convolutions) is the area of projection of the median geniculate body. It is the primary auditory cortex (areas 41 and 42).

Occipital lobe. The visual centers are located in this lobe. There are three areas: (*a*) the striate cortex (17), which receives the lateral geniculate projection; (*b*) the occipital area (18), which surrounds area 17; (*c*) the preoccipital area (19), which extends into the parietal lobe and receives the projection from the pulvinar.

Olfactory cortex. The allocortex (Kölliker's rhinencephalon) is more highly developed in osmatic animals. It occupies the median aspect of the hemispheres. Cortical representation of smell is located mainly in the entorhinal (area 28) and retrosplenic (area 29) regions.

CORTICAL INTEGRATION OF AFFERENT IMPULSES. THE PARIETAL CORTEX

The cortical relay nuclei of the thalamus project to definite areas of the cortex, which have been shown to integrate sensory impulses. Cortical areas corresponding to the great cephalic receptors, *i.e.*, vision (occipital lobe), audition (temporal lobe), olfaction (hippocampus, gyrus cinguli), and taste (postcentral gyrus), will be considered in the respective chapters. Here only the cortical integration of superficial and deep somatic sensations of the body, which are relayed by the posteroventral nuclei to the parietal and frontal cortex, will be considered.

Relatively recent work has demonstrated the existence of two somatic receptor areas. One of these is the well-known postcentral sensory cortex (areas 3-1-2). The other was first reported by Adrian[1] in the cat, but was later found in several other mammals.[2] It is situated in the dorsal wall of the sylvian fissure in the monkey and in a homologous area lateral to the first somatic receiving area and rostral to the auditory area in other animals. Woolsey has proposed that they should be called somatic areas I and II, respectively. Visual and auditory cortical representations are also distributed in two areas, and two receiving and effector areas have been described in the cerebellum (see Chap. 83); therefore "it seems that a dual system of organization may pervade the afferent nervous system."[3]

Stimulation of the parietal cortex. Forty years ago, Cushing made a systematic examination of the parietal cortex, stimulating it electrically in conscious subjects in whom craniotomy had been performed under local anesthesia.

Stimulation of the postcentral gyrus (areas 3-1-2) evokes tactile or pressure sensations, sometimes cold and warmth, but not pain. These sensations are referred to the opposite side of the body. Those provoked by stimulation of the upper part are referred to the lower limb; they are referred to the trunk and upper limb when the middle part is stimulated, and to the neck and face when the lower part is stimulated (Fig. 405).

Stimulation of the precentral gyrus, especially of the rostral lip of the central fissure, provokes tactile sensations and the sensation of movements, although none are performed; occasionally thermal sensations are evoked, but pain is seldom felt. Reference to the periphery is made in a similar way to that observed when the postcentral gyrus is stimulated. Intense stimulation of areas 5 and 7 evokes sensory hallucinations also, which are sometimes visual or auditory. Visual sensations are felt more frequently when the stimulus is applied to more posterior areas, especially area 19. There are no specific areas for touch, thermal, or proprioceptive sensations; all are represented in the same areas. The cortex is not organized according to modalities of sensation, but by dermatomes.

[1] Adrian, E. D., *J. Physiol.*, **100**, 159, 1941.

[2] Woolsey, C. N., *Federation Proc.*, **2**, 55, 1943; **3**, 53, 1944; **5**, 116, 1946; Woolsey, C. N., and D. Fairman, *Surgery*, **9**, 684, 1946.

[3] Woolsey, C. N., *Ann. Rev. Physiol.*, **9**, 525, 1947.

Effects of strychninization of the cortex. The application of strychnine to a limited area of the parietal cortex provokes hyperesthesia, *i.e.*, an increased sensory response to peripheral stimulation, and paresthesia, *i.e.*, abnormal sensations. Disturbances in deep sensibility are always contralateral, while those in cutaneous sensibility are bilateral, but mainly contralateral. Three areas have been differentiated by this method: (*a*) an upper area, corresponding to the lower limb; (*b*) a middle area, corresponding to the trunk and upper limb; (*c*) a lower area, corresponding to the neck and face. These areas extend forward through the depth of the central fissure to the precentral gyrus (areas 4 and 6) and backward to the extreme limit of the parietal lobe. Strychninization of a single point in an area provokes responses corresponding to the whole area, but not to the other two areas; *e.g.*, strychninization of a small part of the arm area is followed by responses in all the upper limb but not in the leg or face. The effects of strychninization of the frontal cortex are not mediated by the parietal cortex; thus, in monkeys, strychninization of the precentral area corresponding to the arm produces its effects after bilateral extirpation of the whole parietal arm area.

Strychninization increases the electrical activity of the cortex; large spikes appear throughout the treated area, but they do not pass to the neighboring areas.[1] The cortex is "set on fire," according to the expression used by Dusser de Barenne, and the fire extends to the corresponding ipsilateral thalamic nuclei. The contralateral thalamus is not disturbed, a fact that is further evidence of the strictly ipsilateral nature of the thalamocortical connections.

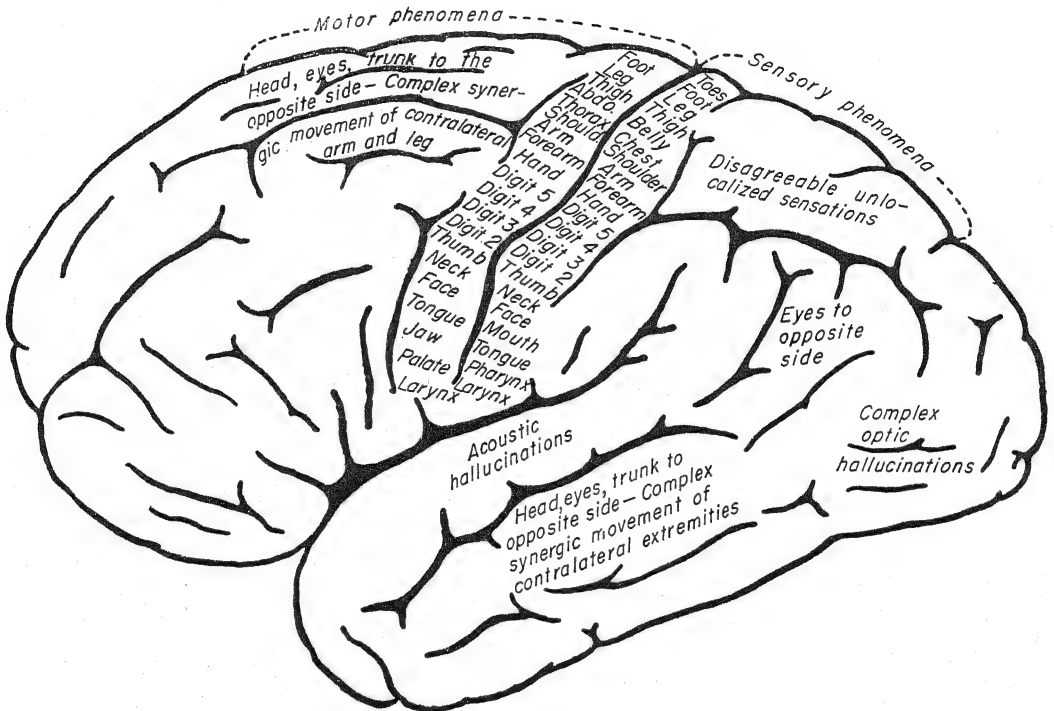



FIG. 405. Motor and sensory localizations on the outer aspect of the cerebral cortex. Effects of stimulation. (*After Foerster.*)

Electrical activity of the cortex evoked by cutaneous stimulation. Stimulation of a limited area of the skin by a tactile stimulus evokes electrical activity (spikes) in a definite point of the contralateral parietal receptor area (areas 3-1-2).[2] If a threshold stimulus is applied, in order to avoid complications due to convergence and spatial summation, activity will

[1] Electrical hyperactivity of the three areas can be provoked by strychninization of area $6\alpha\beta$ in the premotor cortex. Electrical stimulation of this area provokes movements in the whole contralateral half of the body.

[2] WOOLSEY, C. N., W. H. MARSHALL, and P. BARD, *Bull. Johns Hopkins Hosp.*, **119**, 423, 1942.

be seen only in a small area of the cortex; by this method it has been possible to demonstrate that there is a dermatomal representation of the skin surface in the cortex (Fig. 406). Sacral dermatomes are represented in the upper part of the gyrus, and the other dermatomes downward in order of entrance into the spinal cord, except that the area of representation of the upper cervical dermatomes is placed next to that of the upper thoracic segments, and the eighth cervical segment lower down next to the trigeminal area. There is nevertheless a certain degree of overlapping of neighboring dermatomal areas. Representation is mainly contralateral except for the face, which is represented bilaterally. Some of the dermatomes are connected with a wide area of cortex, while others have a much smaller representation. The extent of cortical representation is related to the wealth of peripheral innervation; thus the hand area is large and the trunk area is small, in spite of the lesser area of skin covering the hand. Cortical representation is fixed, *i.e.*, there is no migration of an area from one part of the parietal cortex to another at different times. Therefore it is justifiable to accept the conclusion that the correspondence between skin and cortex is due to fixed fiber pathways.

FIG. 406. Diagrammatic frontal section through postcentral gyrus, showing distribution of dermatomes and sensory localization. Note that C_2 is next to T_1, and C_8 next to the centers corresponding to the face. The trunk has a small area of cortical representation, and the hand and face large areas. *Ophth.*, ophthalmic; *Man.*, mandibular; *Max.*, maxillary; *S.C.M.*, sulcus callosus marginalis. (*Woolsey, C. N., W. H. Marshall, and P. Bard, Bull. Johns Hopkins Hosp., vol. 70, p. 428, 1942.*)

The second somatic receptor area. This area is not distributed into dermatomal areas but is related to the whole surface of the body. It is possible, however, to differentiate three zones, corresponding to the face, arm, and leg. There is always a certain degree of overlapping, but differentiation of the zones is greater in the monkey than in the rabbit. Representation in somatic area II is bilateral, *i.e.*, stimulation of receptors on one side of the body evokes activity in both sides of the cortex, but in the ipsilateral area its amplitude is approximately half that in the contralateral cortex. Apical parts, such as the snout and digits, give an ampler response than other parts, such as the proximal end of a limb, and they are the only parts evoking a response in deeply anesthetized animals. In somatic area I richness of innervation is reflected by a large area of cortical representation; in somatic area II it is amplitude of response that increases with density of innervation.

Somatic area II is not dependent on somatic

area I; they each have independent afferent paths, and one of them can respond when the other has been rendered inactive. Area II is related bilaterally to the paramedian receptor-effector area of the cerebellum.

The functional significance of somatic area II is not yet well understood. Perhaps it plays a part in the integration of complex and widespread sensations, while area I is the site of the mechanism for precise localization and discrimination.

Destructive lesions of the parietal lobe. Destructive lesions of the parietal lobes, provoked experimentally or due to trauma or disease, cause sensory disturbances in the contralateral half of the body. Temperature and pain sensations show little or no change. Tactile sensibility shows the loss of precise localization and discrimination, and its differential thresholds are raised (*e.g.*, for two-point sensibility). Proprioceptive sensibility, *i.e.*, the sensation of active and passive posture and movement, is considerably disturbed, and maximum losses are observed in complex sensation patterns such as stereognosis. The patient responds irregularly and is easily fatigued when the side opposite to the lesion is stimulated, but responds normally to ipsilateral stimulation. According to Head's concept of the mechanism of sensation, epicritic sensibility is lost.

Experimental removal of the parietal cortex has been combined with the methods of analysis used by experimental psychologists, which have been applied to patients with cerebral lesions. Thus it has been possible to demonstrate that certain behavior patterns require cortical integration of somatic sensibility. The placing and hopping reactions (see Chap. 81) and responses learned by training to differentiate weights, or rough from smooth surfaces, or geometric shapes (stereognosis) are impaired by removal of the parietal cortex. The deficit increases according to the degree of "encephalization" of the species, *i.e.*, it is more pronounced in man and primates than in lower animals. There is a certain amount of recovery of function as time passes, especially if part of the parietal lobe remains. The hopping reactions and weight differentiation are recovered sooner and more completely than placing reactions and stereognosis.

Different methods of exploration in some cases apparently give contradictory results, but this is not really the case; each method shows a particular aspect of cortical function that is not made evident by other methods. Thus strychninization increases excitability and, by tending to spread the excitatory state, reveals the extent of interconnections between different centers, masking the finer aspects of these connections. On the other hand, stimulation of receptors and registration of electrical activity in paths and centers permits the exploration of more precise and intimate connections between the receptors and centers. This method reveals focal representation that is blurred by strychninization. Removal of parts of the cortex shows how in the integration of complex responses the whole cortex plays a part, and how one area may to a certain extent replace others.

BIBLIOGRAPHY

BARD, P., Studies on the Cortical Representation of Somatic Sensibility, *Harvey Lect.*, 1938, pp. 143–169.

DÉJERINE, J., "Séméiologie des affections du système nerveux," Masson et Cie, Paris, 1914.

FULTON, J. F., "The Physiology of the Nervous System," 3d ed., Oxford, New York, 1949.

HEAD, H., Sensation and the Cerebral Cortex, *Brain*, **41**, 57, 1918 (*Studies in Neurology*, **2**, 716, 1920).

HINES, M., The Somatic Function of the Nervous System, *Ann. Rev. Physiol.*, **14**, 391, 1952.

KAADA, B. R., Electrical Activity of the Brain, *Ann. Rev. Physiol.*, **15**, 39, 1953.

WALKER, A. E., "The Primate Thalamus," University of Chicago Press, Chicago, 1938.

CHAPTER 75

Chemical Senses. Smell and Taste

CERTAIN SUBSTANCES in solution stimulate special receptors and activate neurophysiologic mechanisms known as the chemical senses. Two of these, smell and taste, have differentiated receptors, but there are also free nerve endings in the nasal (nonolfactory), oral (nongustatory), respiratory, anal, and genital (vulva and glans) mucosae which can be stimulated by dissolved substances, giving rise to what is known as the general chemical sense.

THE GENERAL CHEMICAL SENSE

The skin of aquatic vertebrates (fishes and amphibians) is sensitive to chemical stimuli; acids placed on the skin of the frog or toad provoke remarkable defensive reactions. Other vertebrates develop a horny layer on the skin surface which protects them from the action of chemical agents, but if this layer is removed the free nerve endings can be stimulated by substances in solution. The mucosae that are in contact with the external environment (*i.e.*, those just mentioned) are less well protected and therefore respond to chemical stimuli. The general chemical sense forms part of the pain sense; it has the same receptors and afferent nerve paths and evokes painful sensations.

There are chemical receptors connected with visceral nerves; thus the aortic and carotid bodies are sensitive to variations in the partial pressures of O_2 and CO_2 and in hydrogen ion concentration, and to certain drugs. Stimulation of these receptors provokes the discharge of impulses along the nerves of de Cyon and Hering, which initiate reflexes regulating the circulation and respiration. Conscious sensations are not evoked by the stimulation of visceral chemoreceptors (see "Vascular reflexes," p. 195, and "Chemoreceptors," p. 296).

A given substance may stimulate the mucosae and the differentiated receptors, but each receptor needs a particular concentration (strength of stimulus); *e.g.*, the majority of people can detect ethyl alcohol by smell in a concentration of approximately 0.000125 *M*, but a 3-*M* solution is necessary to stimulate taste and a 5- to 10-*M* solution to excite the mucosae. The intact skin is not sensitive to alcohol, but if the horny layer is removed the nerve endings are made accessible and a smarting sensation is produced. Some substances stimulate only one type of receptor; *e.g.*, mercaptan has no taste, but it is a powerful stimulant of smell.

SMELL

Receptors. In the upper part of each nasal cavity there is an area of about 250 sq. mm. on the septum and the superior concha, where the nasal mucosa has a yellow color; this is the olfactory epithelium. This epithelium is made up of three kinds of cells: (*a*) *basal cells*, which are small, conic elements distributed in a single layer on the chorion; (*b*) *sustaining cells*, which are elongated and have a cuticle on the free surface joined to that of other cells so as to form a continuous membrane with perforations through which the cilia of the olfactory cells pass out into the nasal cavity; (*c*) *olfactory cells*, which are modified neurons distributed regularly among the sustaining cells.

The olfactory cells have a proximal and a distal process. The latter is a thick dendrite ending in 6 to 8 fine cilia of 2-μ length, which perforate the membrane formed by the sustaining cells and emerge on the surface of the mucosa. The proximal process is a slender axon 1 μ in diameter; this unmyelinated fiber passes through the cribriform plate of the ethmoid and ends synaptically on the mitral and tufted cells

of the olfactory bulb. Each one of the bulb neurons receives the fibers of several olfactory cells. Extirpation of the olfactory bulb causes retrograde degeneration of the olfactory cells in the mucosa which is complete within a week. Serial sections of the whole olfactory mucosa after ablation of different parts of the bulb in the rabbit show that the degenerated cells are distributed in a manner that indicates there is some degree of regional projection of the olfactory epithelium onto the bulb.[1] Topographic organization is not, however, as precise as in other sense organs, *e.g.*, the almost point-to-point projection of the fovea of the retina on the lateral geniculate body. Electrical methods have also shown there is some degree of localization in the bulb.[2] The olfactory cells are at the same time receptors and conductors; they are the first-order neurons of the olfactory path. This is the only sensory apparatus in which such a primitive structure, similar to that of invertebrates, is found.

Stimulation of the Olfactory Receptors

The olfactory cells are stimulated by small particles given off by aromatic substances. The particles are conveyed into the nasal cavity in the inspired air and by diffusion reach the olfactory area, where they are dissolved in the fluid secreted by Bowman's glands. These glands are situated under the olfactory epithelium, and they have a short duct opening on the surface of this epithelium. The dissolved odoriferous substances act on the cilia of the olfactory cells. The membrane of these cells has a lipoprotein structure; therefore the odoriferous potency of a substance is conditioned by its solubility in fats. The odorous particles are relatively heavy; they diffuse slowly, and the latency of the sensation is consequently lengthened. Stimulation by the first particles to arrive provokes an olfactory reflex—a short, sharp inspiration (a sniff), sometimes accompanied by extension of the neck, which directs the inspired air into the upper part of the nasal cavities, causing a greater number of particles to reach the sensory area. The weight of the particles makes them fall and lie on the ground, so that dogs and other animals with a well-developed sense of smell can follow a trail guided by the odor left by the prey.

Olfactory cells in fishes are sensitive to mechanical stimulation (gentle stroking). In mammals they seem to be sensitive to pressure; there is a burst of impulses at each inspiration, even when the inhaled air is odorless.

Olfactory thresholds. Several methods have been used to determine the threshold of sensation. The air inhaled must contain no odor except the one being tested, and it must be delivered at constant pressure. The *absolute threshold* is obtained by evaporating known quantities of the odorous substance in a given volume of air; the lowest concentration that can be detected by smell is the threshold. The *relative threshold* can be determined by means of Zwaardemaker's olfactometer. This instrument consists of a tube, one end of which is introduced into the nostril and the other into a second tube of greater diameter. The substance tested is spread on the porous inner surface of the outer tube. The air going into the nose must pass through this tube and is thus loaded with aromatic particles. The amount of these is proportional to the length of tube through which the air has passed, and it is measured in arbitrary units of length called olfacties. A double tube, one for each nostril, permits the combination of different substances by simultaneous application. Elsberg[1] has designed an olfactometer that pumps air saturated with the odorous substance into the nostril while the subject holds his breath. The force of the blast is controlled so that it is always the same. The threshold is the minimal identifiable odor, *i.e.*, the smallest amount of saturated air that provokes a sensation of smell.

Olfactory stimuli. The thresholds of different aromatic substances differ widely. Artificial musk is one of the most potent stimulants of smell; it can be perceived in a concentration of 0.04 μg per liter. Natural musk must be in a concentration 1,000 times higher to provoke a sensation, and ethyl ester can be detected only if there is 5 or 6 mg. per liter of air. The stimulating power of a substance is related to its boiling point, because the odoriferous particles are given off by evaporation. The more volatile substances are stronger stimuli than less volatile substances. Thus in the alcohol series the boiling point decreases and the odoriferous power in-

[1] Le Gros Clark, W. E., *Nature, London,* **165,** 452, 1950.

[2] Adrian, E. D., "Sensory Integration," University Press, Liverpool, 1949.

[1] Elsberg, C. A., in Glasser, "Medical Physics," Year Book Publishers, Inc., Chicago, 1944.

creases as the carbon chain lengthens; methyl alcohol produces the threshold sensation at a concentration of 1,000 μg per liter of air, ethyl alcohol at 250 μg, propyl alcohol at 5 to 10 μg, butyl alcohol at 1 μg, and isoamyl alcohol at 0.1 μg.

Classification of odors. Aromatic substances produce different sensations; they do not all have the same smell. Many attempts have been made to classify odors. Zwaardemaker's and Henning's classifications are among the best known, but none is satisfactory. It has not been possible to establish a definite relation between the chemical structure of a substance and its odoriferous properties. The position of the chemical groups OH, CH_2, NH_2, etc., on the ring or side chain of a substance in the cyclic series has a great importance in the determination of its odorific as well as other pharmacodynamic properties. In some cases substances with similar chemical structures have very different smells, while in others two substances with widely different chemical structures have the same or a similar smell.

Sensibility of the receptor. The threshold for a given substance varies within relatively narrow limits in different individuals, and it is fairly constant in the same subject. In women sensitiveness to smell increases before and during menstruation. Obstruction of the nasal cavities due to malformation, congestion of the mucosa, or inflammatory processes (*e.g.*, the common cold) raises the threshold by hindering the passage of odoriferous particles to the olfactory area. The olfactory epithelium is very sensitive to harmful agents, and once the olfactory cells are destroyed they are not replaced.

Difference thresholds. The capacity to distinguish one smell from another, or difference in the intensity of smells, is not easily determined because of the rapidity of adaptation and fatigue. It varies considerably in different subjects. Training is of great importance, but there are also inborn factors that condition the capacity to discriminate smells.

Localization. Olfactory sensations are not referred to the receptor or to any definite point of the environment. The origin of a smell can be localized only with the aid of other sensations.

Adaptation and fatigue. A given stimulus soon loses its efficiency on the olfactory receptor, especially if it is a strong one. This is due in part to adaptation (see Chap. 71) and in part to fatigue. Recovery, however, takes place very rapidly. Fatigue produced by one substance diminishes, but does not suppress, the effects of others of different odor. If fatigue is marked, there is a transitory partial anosmia for the substances that have produced fatigue.

Combination of stimuli. When several stimuli are applied simultaneously, at first one of them can predominate, but after a time, owing to adaptation and fatigue, it loses predominance and gives place to another stimulus. Sometimes the smell of one substance neutralizes that of another; thus by means of Zwaardemaker's double olfactometer it is possible to demonstrate that the smells of rubber and paraffin neutralize each other when applied in the proportion of 14:8.5. Some of the deodorizing substances act by either predominance or neutralization. Simultaneous application of two substances of similar strength but of different smell can produce a sensation completely different from that of either. The art of perfumery consists in finding novel and agreeable combinations. Frequently these are obtained by adding very small amounts of substances which in high concentration have a repugnant smell.

All the olfactory cells are apparently the same. There is no evidence that certain cells are specifically sensitive to odoriferous substances of a given chemical structure. Adrian has shown that the anterior part of the olfactory mucosa is more apt to be stimulated by substances soluble in water, and that impulses from this area are discharged with a shorter latency. The posterior part of the mucosa is stimulated preferentially by substances soluble in oil, and there is a longer latency in the discharge of impulses. "Discrimination depends in the main on differences in the sensitivity of different receptors to different kinds of smell, for when we use threshold concentrations we often find impulses are set up in different conducting units. Most of them, however, react to most smells when their concentration is much above threshold."[1] Different smells would, therefore, be due to the projection on the cortex of different patterns of excitation, arising not only in the olfactory receptors, but also in the nasal mucosa innervated by the fifth cranial nerve, *e.g.*, in the case of ammonia.

Enzymes (phosphatases, nucleotidases, esterases, etc.) have been found in the olfactory mucosa and in the mucous and serous secretion

[1] ADRIAN, E. D., *Nature, London*, **168**, 1021, 1951.

of Bowman's glands. They have been considered to have a part in the stimulation of olfactory receptors.[1]

Olfactory Pathway and Centers

Magendie and other physiologists in the nineteenth century believed that the trigeminal was the olfactory nerve. This is an erroneous conception, as is proved by the fact that extirpation of the gasserian ganglion does not abolish the sense of smell. The trigeminal nerve fibers ending in the olfactory area convey painful impulses. Congenital atrophy of the olfactory bulb, on the other hand, causes permanent and complete anosmia. In recent years several physiological methods have been used in the study of the olfactory paths and centers: (*a*) stimulation of the receptor or of the afferent paths by odorous substances or electrical stimuli, combined with the registration of the spread of the action potentials; (*b*) establishment of olfactory conditioned reflexes; (*c*) the discrimination method, in which animals are trained to distinguish a box with a given smell, containing food, from another, with a different smell, which has no food.

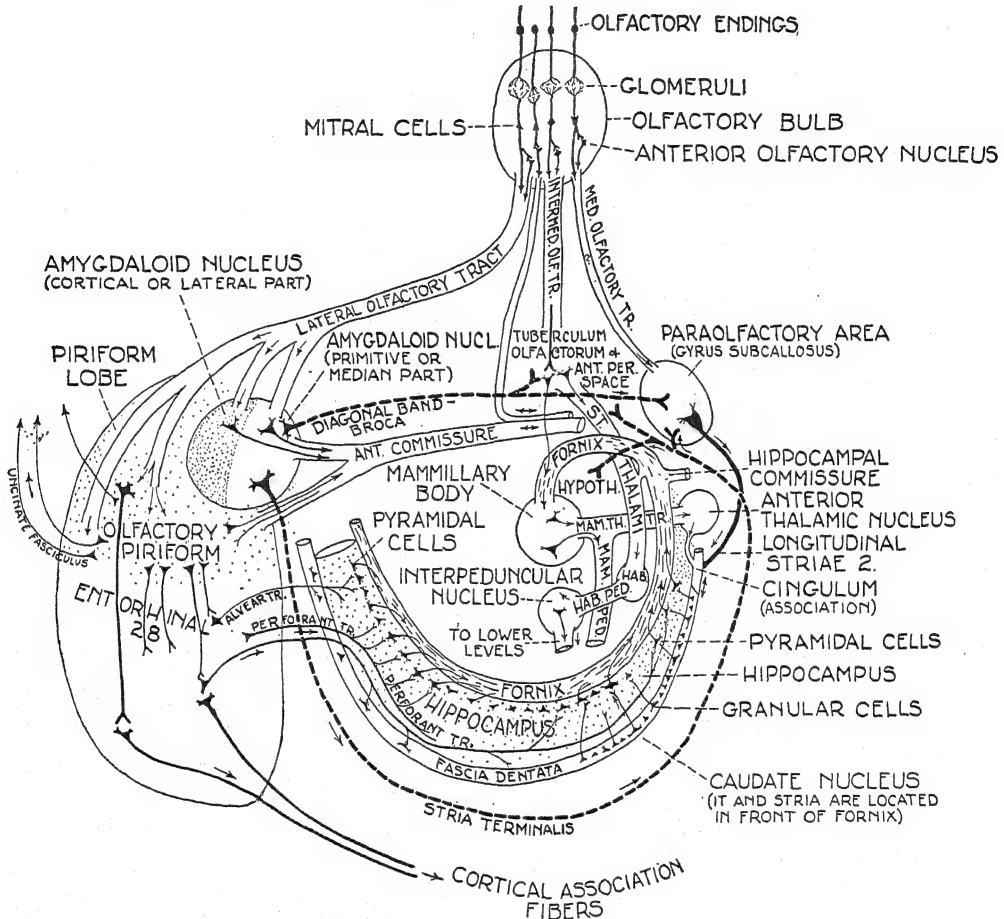



Fig. 407. Diagram of olfactory system (left side, from above). (*Allen, W. F., J. Comp. Neurol. vol. 88, p. 425, 1948.*)

The rhinencephalon or olfactory brain (Fig. 407) is greatly developed in fishes; in selachians it forms almost the whole of the forebrain. The development of the neopallium (the cerebral cortex corresponding to somatic sensation and motility) diminishes the importance of the rhinencephalon (archipallium) in mammals. Nevertheless in macrosmatic species, such as the rodents, in which olfaction plays an important

[1] Baradi, F. A., and G. H. Bourne, *Nature, London*, **168**, 977, 1951; *Science*, **113**, 660, 1951; Kistiakowsky, G. B., *Science*, **112**, 154, 1950.

part, the rhinencephalon is relatively large. Man is a microsmatic animal, and the olfactory cortex is therefore proportionately small.

The olfactory bulb is a diverticulum of the rhinencephalon, with a laminated structure similar to that of the rest of the archipallium. The central cavity has been filled by neuroglia, around which there are myelinated fibers, the axons of the mitral and tufted cells that form the olfactory tract. There are three kinds of neurons in the bulb: the mitral and tufted cells, which send their dendrites outward to the periphery, where they are joined synaptically to the unmyelinated fibers of the olfactory cells; and the deeply placed granules, which send short axons toward the surface of the bulb.

The olfactory tract is covered dorsally by a thin layer of gray matter; it is, therefore, a rudimentary gyrus. It lies on the orbital aspect of the frontal lobe and ends posteriorly in the olfactory tubercle, dividing up into three branches: (*a*) the *medial* olfactory tract, or gyrus, which merges into the paraolfactory area, communicating with the cingulum through the longitudinal striae; (*b*) the *lateral* olfactory tract, or gyrus, which passes into the amygdaloid nucleus and the uncinate gyrus (this anterior part of the hippocampus, together with the lateral tract, is known as the pyriform area); (*c*) the *intermediate* tract or gyrus, which is less distinct and passes directly into the anterior perforated space.

The olfactory cortex is made up of the hippocampus, the gyrus dentatus, the cingular gyrus, and the anterior perforated space. It forms an almost complete circle around the corpus callosum. The olfactory cortex receives fibers mainly from the amygdaloid nucleus and the pyriform area. The afferent olfactory pathway, unlike other sensory paths, does not pass through the thalamus; there is no primary thalamic projection, and no thalamocortical olfactory fibers are known. On the other hand the principal outgoing tracts from the olfactory cortex form the fornix, which goes from the hippocampus to the mammillary body of the hypothalamus. The mammillary body sends out fibers to the anterior nuclei of the thalamus (mammillothalamic tract) and the brain stem (mammillopeduncular tract). The anterior perforated space sends fibers to the habenula, which emits fibers to the mesencephalic nuclei. The anterior and the hippocampal commissures establish connections between both hemispheres. The olfactory centers are extensively connected with other cortical regions.

Stimulation of the olfactory epithelium with aromatic substances induces electrical activity in the olfactory tracts and the hippocampus. This activity is abolished if the olfactory epithelium is anesthetized. Fox, McKinley, and Magoun[1] have used a Horsley-Clark apparatus to place the electrode accurately in the deep brain structures and pyriform areas of the cat. They have carefully explored olfactory nerve centers in this animal and have found that stimulation of the olfactory bulb provokes electrical activity in the anterior perforated substance and the prepyriform and pyriform areas. They also observed spread of electrical activity from the olfactory areas to the putamen and globus pallidus, suggesting the existence of connections between the olfactory centers and the striatum. Stimulation of the pyriform area evokes electrical potentials in the prefrontal area, thus showing a connection between these centers.

If all the cortex except the olfactory centers, pyriform area, and anterior perforated space is removed in monkeys and cats, and the olfactory centers are stimulated electrically several months later, reactions are observed similar to those provoked by the smelling and ingestion of food, *i.e.*, sniffing, secretion of saliva, chewing, movements of the tongue, swallowing, and in monkeys the emission of guttural sounds.[2]

Allen,[3] in a remarkable series of studies, has employed the conditioned-reflex technique, associated with localized extirpation of certain areas. Conditioned reflexes to aromatic substances are not established, or are suppressed, if the olfactory bulb or the olfactory tract is removed. If these substances also have an irritating effect on the nasal mucosa (ether, chloroform, eucalyptus, camphor, etc.), reflexes can be established through the trigeminal nerve.[4] Large parts of the occipital, parietal, and tem-

[1] Fox, C. A., W. A. McKinley, and W. Magoun, *J. Neurophysiol.*, **7**, 1, 1944.

[2] Rioch, D. McK., and C. Brenner, *J. Comp. Neurol.*, **68**, 495, 1938; Smith, W. K., *Am. J. Physiol.*, **133**, 451, 1941.

[3] Allen, W. F., *Am. J. Physiol.*, **118**, 532, 1937; **121**, 657, 1938; **126**, 419, 1939; **128**, 754, 1940; **132**, 81, 1941.

[4] This would explain some old observations of cases of atrophy of the olfactory bulb, in which the sense of smell was reported as still present (Bernard, C., "Leçons sur le système nerveux," p. 229; Testut, "Anatomie humaine," 6th ed., vol. 3).

poral lobes and 90 to 100 per cent of the hippocampus of both sides can be removed in the dog without impairing the establishment or conservation of positive and negative olfactory conditioned reflexes and discrimination. Bilateral extirpation of the pyriform lobes and amygdaloid nuclei or of the prefrontal cortex hinders the acquisition of conditioned reflexes and especially of olfactory discrimination. None of these operations prevents the animals from recognizing meat when blindfolded. Allen[1] suggests that Ammon's horn functions as a mechanism for the amplification of simple olfactory reflexes.

The rhinencephalon does not have exclusively olfactory functions. Stimulation of the limbic lobe provokes respiratory, cardiovascular, and other visceral responses, *e.g.*, contraction of the bladder. Stimulation of the anterior cingulate gyrus has an inhibitory effect, although no paths from this area to the reticular formation have been described. Auditory and optic stimulation evokes electrical potentials in the cingulate gyrus and the hippocampus; the latter also receives somesthetic impulses[2] and intercortical fibers from areas 4, 6, 8, 9, and 10.

The rhinencephalon also takes part in the integration of emotional response and general behavior. Bard and Mountcastle[3] have shown that hyperirritability and sham rage, which appear in cats after removal of the greater part of the telencephalon, are not observed after extirpation of the neocortex if the rhinencephalon remains undamaged; on the contrary, the animals are remarkably placid.

Physiologic Significance

Smell is of fundamental importance in the search for food in many species. In man it has little significance in this respect, but it plays a part in digestion, as the smell of food stimulates salivary, gastric, and other digestive secretions and activates gastrointestinal movements. Smell is important in the initiation of sexual reflexes. In some animals it is an auxiliary in defense, as it gives warning of the proximity of an enemy. Olfactory sensations have a high affective tone of pleasure or displeasure; as a rule any strong smell is disagreeable. They also have a great capacity for awakening memory; possibly this is due to the numerous connections with other parts of the cortex. Through these connections the olfactory centers have an important nonspecific influence upon other cortical and subcortical areas, lowering the threshold and thus facilitating many reactions, or else increasing inhibition. In states of depression it is said that one has lost the taste for life; from a neurophysiological standpoint, it would perhaps be more appropriate to say that life has lost its perfume.

[1] Allen, W. F., *J. Comp. Neurol.*, **88**, 425, 1948.

[2] Glees, P., C. W. M. Whitty, and H. Cairns, *J. Neurol., Neurosurg. & Psychiat.*, **13**, 178, 1950.

[3] Bard, P. E., and V. B. Mountcastle, *Research Publ., A. Nerv. & Ment. Dis.*, **27**, 362, 1947.

TASTE

What is commonly called "taste" is a complex sensory pattern made up not only of sensations arising in the specific taste receptors but also of those arising in general chemical, tactile, warm and cold receptors in the mouth, and especially of olfactory sensations. If the latter are suppressed, taste, and particularly taste discrimination, is considerably diminished. The gustatory sense only will be considered here.

Receptors. Taste receptors are distributed on the dorsal aspect of the tongue, the laryngeal aspect of the epiglottis, the pharynx, the soft palate (except the uvula), the anterior faucial pillars, and a few on the cheeks and the undersurface of the tongue. The taste organs appear in the course of the third month of fetal life. In the child they are not so widespread as in the fetus, but they are still found over the whole upper surface of the tongue and the mucosa of the cheeks. In the adult they diminish with age, first those on the middorsal region and then those on the tip of the tongue. They are mostly found in the fungiform, foliate, and circumvallate papillae. There are no taste buds in the filiform papillae. The small serous salivary glands that pour their secretion into the papillae are an important part of the mechanism of taste. They first dissolve sapid substances, thus facilitating their action on the receptors, and then wash them out, leaving the receptors in condition to receive another stimulus. The taste buds are made up of two types of cells, the sustentacular and the gustatory; the latter are spindle-shaped, and the peripheral end is drawn out into a ciliary process or gustatory hair, which projects through the pore of the bud. Both types of cells are innervated.

Thresholds. The taste receptors are stimulated by introducing into the mouth a certain

amount of the sapid substance in a solution of known concentration. Localized stimulation is provoked by placing a drop of the solution on the area examined, taking care not to let it spread. The thresholds vary considerably from one subject to another, but are fairly stable in the same subject. The threshold varies on the different areas of the tongue according to the distribution of the taste buds. In animals the thresholds can be determined by the conditioned-reflex method, the discrimination method, or the preference method. In the last method, the subject is given the choice of drinking tap water or a solution of the sapid substance; the minimum concentration that shows a differential consumption with water—either more, if the taste is preferred, or less, if it is rejected—gives the preference threshold, which is not equivalent to the absolute threshold. Distilled water at pH 7 is often considered the standard insipid substance, yet it has a peculiar taste similar to that of dilute alkali, owing to the fact that it has no CO_2. Saliva is the standard by which other tastes are judged (Moncrieff). An electric stimulus applied to the tongue provokes electrolysis, and usually a sour taste is perceived. The registration of action currents of the sensory nerves innervating the taste buds when these are stimulated has also been used to study the effect of different substances.

Taste receptors can also be stimulated by intravenous injection of a substance; *e.g.*, if "decholin," the sodium salt of dehydrocholic acid, is injected into a vein, a bitter taste is perceived when it arrives at the taste buds. This fact has been used to measure the circulation time.

Classification of tastes. Four tastes are recognized by the majority of normal human subjects: (*a*) salt; (*b*) acid or sour; (*c*) sweet; (*d*) bitter.

Salt taste. Dissociated salts stimulate the taste buds, but stimulation by nondissociated salts cannot be ruled out. The taste of common salt (NaCl) is the typical salt taste. Both the anion and the cation play a part in evoking this taste. Cl^-, Br^-, I^-, $SO_4^{=}$, NO_3^- have a saline taste, but all their salts do not taste the same. Thus NaBr has a salty taste, KBr tastes salt and bitter, and RbBr tastes bitter. Salts of cations of low molecular weight Na^+, Li^+, NH_4^+, taste salty, K^+ salts taste both salt and bitter, and salts of heavy cations such as Ca^{++}, Mg^{++}, Rb, and Cs taste bitter.

The absolute threshold of the saline taste of NaCl is given by a 0.087 per cent solution; some subjects perceive a saline taste in 0.02 per cent NaCl. If NaCl solutions are compared with distilled water, the salt-taste threshold is at 0.01 per cent, or even less (0.007 per cent).[1]

Sour taste. This is a property of all acids and is mainly dependent on the H^+ concentration. Mineral acids, such as HCl, H_2SO_4, H_2NO_3, etc., cannot be distinguished by their taste, and the strength of the sensation evoked is proportional to the H^+ concentration. The threshold is about pH 3.4 to 3.5. Organic acids, such as acetic, butyric, and citric, have a higher threshold if the minimum concentration of the acid is considered; *e.g.*, the threshold for HCl is 0.00125 *N* solution, and for acetic acid it is 0.005 *N*. Organic acids are weak acids, and less dissociated than mineral acids; in the example just given, the HCl solution has between four and five times the H^+ concentration of the acetic acid solution. Therefore organic acids are more efficient stimuli of taste buds than mineral acids. The threshold for organic acids is pH 3.7 to 3.9. Moreover, buffer solutions of weak acids evoke a sour taste at a much lower H^+ concentration, *e.g.*, the threshold for a buffer solution of acetic acid and sodium acetate is pH 5.6.[2] The undissociated salt plays a part in this case. The greater efficiency of organic acids as stimulators of taste has been attributed to their higher solubility in lipids, and it is supposed they can penetrate into the taste cells and increase intracellular H^+ concentration more rapidly than mineral acids.[3]

Sweet taste. Polyatomic alcohols, *i.e.*, substances which have several CH_2OH groups, have a sweet taste. The most common are glycerol, and sugars, such as hexoses (glucose, fructose, etc.) and their polymers (sucrose, lactose, etc.). The sweet taste is also a property of substances with a completely different chemical structure, *e.g.*, α-amino acids, lead acetate, chloroform, beryllium salts, saccharin (*o*-sulphobenzimide), and dulcin (*p*-ethoxy-phenylurea).

The threshold for the sweet taste of sucrose is given by 0.4 to 0.5 per cent solutions. The sweetness of a substance is usually compared with that of sucrose (Table 97). With the ex-

[1] RICHTER, C. P., and A. MACLEAN, *Am. J. Physiol.*, **126**, 1, 1939.

[2] LILJESTRAND, G., *Arch. néerl. de physiol.*, **7**, 537, 1922.

[3] TAYLOR, N. W., *J. Gen. Physiol.*, **11**, 207, 1928.

ception of fructose, most of the sugars are not so sweet as sucrose, but there are synthetic substances which are several hundred times sweeter. The relative sweetness of sugars varies with the concentration. Isosweet curves of several sugars (glucose, galactose, lactose, etc.) compared with sucrose are not lineal but exponential.[1]

Table 97. Relative Sweetness of Different Substances

Substance	Relative sweetness
Sucrose	1.0
Fructose	1.73
Glycerol	1.08
Glucose	0.74
Galactose	0.32
Maltose	0.32
Lactose	0.16
Dulcin	265
Saccharin	675
4-nitro-2-aminophenylpropyl ether	3,300

Sources: BIESTER, WOOD, and WAHLIN, *Am. J. Physiol.*, **73**, 387, 1925; GILMAN and HEWLETT, *Iowa State Coll. J. Sc.*, **4**, 27, 1929; BLANKSMA and VAN DER WEYDEN, *Rec. de trav. chim d. Pays-Bas*, **59**, 629, 1940.

Bitter taste. Alkaloids (quinine, strychnine' etc.), glucosides, salts with a high molecular weight (Ca, Mg, Rb, Cs salts), and other substances such as picric acid have a bitter taste. The threshold for quinine is given by a 0.0016 per cent solution; strychnine is about three times as bitter as quinine and brucine (dimethoxystrychnine) more than ten times as bitter.

Certain substances have more than one taste, *e.g.*, the salts with saline and bitter taste. Others, such as the glucoside dulcamarin, taste sweet and bitter; pyridine 2,5-dicarboxylic acid is sweet, bitter, and sour.

Solubility in water is an essential prerequisite for a substance to have taste. Solubility in lipids may also play a part in enhancing this property. Chemical structure is another essential prerequisite. This is not so apparent in the more primitive salt and sour tastes, in which acids and salts taste much alike; but it is of fundamental importance for the sweet and bitter tastes, which, according to Moncrieff, are a later and higher specialization. The capacity to stimulate a taste receptor is due to special chemical groups (saprophores), the loss or modification of which renders the substance tasteless or gives it another taste. Thus *p*-etoxyphenylurea (dulcin) is extremely sweet and *p*-etoxyphenylthiourea is very bitter; dimethoxystrychnine is three to four times more bitter than strychnine.

[1] CAMERON, A. T., *Canad. J. Research*, **22**, 45, 1944; **23**, 139, 1945.

Distribution of the taste receptors. A taste bud is apparently stimulated exclusively by the substances in one of the four groups mentioned, but in the same papilla there can be buds corresponding to two or more tastes. In man, the salt taste receptors are more numerous in the tip and the anterior part of the lateral margins of the tongue, acid in the lateral margins, sweet in the tip, and bitter in the base, connected with the large circumvallate papillae. A given substance can stimulate two types of organs. Thus $MgSO_4$ tastes salt on the tip and bitter on the base of the tongue; hexamine is sweet and also bitter.

In the cat,[1] three types of taste organs have been discovered by registering the action potential in the chorda tympani after reducing it by dissection to only a few fibers: (*a*) those stimulated by acids; (*b*) those stimulated by acids and salts (NaCl); (*c*) those stimulated by acids and bitters (quinine). Sweet substances apparently do not stimulate any receptors. These experiments show that in the cat either a receptor is stimulated by substances of two tastes or else the final branches of one axon end in receptors of two different types; in the second case the sensory unit would not be homogeneous. The nerve impulses are the same whatever stimulus is used, a fact that proves conclusively that the quality of sensation is not dependent on differences in the nerve impulses. In this species salt receptors are more numerous in the anterior part, and bitter ones in the back, while acid receptors are distributed all over the tongue but are relatively scarce on the middorsal aspect. In the dog there are receptors for the four tastes. Sweet receptors are fairly widely distributed; they are stimulated by sucrose, glycerol, lead acetate, and other substances tasting sweet to man, but not by saccharine. Taste impulses are conducted by relatively fine fibers, 4 to 6 μ in diameter in the dog (calculated from the spike potential). Impulses arising in receptors stimulated by strychnine and other bitter-tasting substances are conducted in very fine fibers.[2] The initial frequency of the impulses depends on the

[1] ZOTTERMAN, Y., *Skandinav. Arch. f. Physiol.*, **72**, 73, 1945.

[2] ANDERSSON, B., S. LANDGREN, L. OLSSON, and Y. ZOTTERMAN, *Acta physiol. Scandinav.*, **21**, 105, 1951.

concentration of the sapid substance, *i.e.*, the strength of the stimulus. Adaptation is relatively slow. There are no morphologic differences between the taste buds sensitive to different kinds of sapid substances. Sensitiveness to different substances is probably due to differences in the chemical constitution of the receptors. Localized stimulation has shown that there are individual receptors that respond exclusively to one of the four types of stimuli. Pfaffmann's[1] experiments in the cat, recording the action potentials of single nerve fibers, also demonstrate the existence of receptors that respond to only one type of sapid substance. Moreover, cocaine applied on the tongue abolishes first the sensation of bitter, then sweet, then salt, and finally acid. Sweet and bitter tastes are selectively abolished by the local application of gymnemic acid, which leaves salt and sour unimpaired.

Taste sensibility. There are considerable individual differences in taste sensitiveness. Certain subjects cannot perceive any taste; they are "taste-blind," a condition known as ageusia. Loss of one taste, *e.g.*, sweet or bitter, leaving the others unimpaired has not been observed. The majority of individuals recognize as sweet a solution of 0.41 per cent of saccharose; for others even a 10 per cent solution is tasteless. Ethyl alcohol has no taste for some subjects, who drink it in concentrations of 50 per cent and more, while most subjects find it disagreeable at 10 per cent or even less.

Certain substances which taste bitter to the majority of individuals are tasteless to others, *e.g.*, *p*-ethoxyphenylthiourea. This substance and others, such as phenylthiocarbamide (PTC) and its derivates, which also have a bitter taste, have been widely used to investigate taste deficiencies. The threshold for PTC varies considerably; the minimum concentration that can be distinguished from water is 0.000005 per cent for some individuals and 0.1 per cent for others, with a maximum frequency at 0.0003 per cent.[2] A certain number of subjects (3 to 5 per cent) are practically "taste-blind" to this substance. In women the threshold was found to be slightly lower than in men, and when the response to 0.05 per cent PTC was tested, 22.2 per cent

[1] PFAFFMANN, C., *J. Cell. & Comp. Physiol.*, **17**, 243, 1941.

[2] SALMON, T. N., and A. F. BLACKSLEE, *Proc. Nat. Acad. Sc.*, **21**, 78, 1935; RICHTER, C. F., and K. H. CLISBY, *Am. J. Physiol.*, **134**, 157, 1941.

were found to be nontasters; the corresponding figure for men being 25.9 per cent.[1] This type of "taste-blindness" is inherited according to the mendelian laws. When both parents are nontasters, the offspring are all nontasters.[2]

Diversity in taste has been remarkably demonstrated by the following experiment: At a meeting of scientists (*i.e.*, trained observers experienced in the objective consideration of phenomena), the taste of an unknown substance (maltose) was tested. The answers obtained were (*a*) sweet; (*b*) bitter; (*c*) acid; (*d*) salt; (*e*) insipid; (*f*) complex—that is to say, all the possible answers. The old Latin tag *ae gustibus non est disputandum* is apparently true in a literal as well as a figurative sense.

Taste threshold varies with the temperature of the solution used as a stimulus. It is at a minimum between 20 and 30°C. and rises at temperatures below and above these. Changes in the concentration in the blood, and therefore in the taste buds, of salts modify the preference threshold for these substances. Thus in adrenal insufficiency there is a decrease in blood sodium, and a specific appetite for sodium is awakened. Normal rats recognize and prefer a 1:2,000 NaCl solution; adrenalectomized rats distinguish and prefer a solution of 1:33,000 NaCl.[3] Adrenalectomized rats spontaneously increase their NaCl consumption, compensating for the excess loss of salt in the urine. Specific appetite for salt has also been observed in patients with Addison's disease (adrenal insufficiency). An increase in blood sodium causes the specific appetite for sodium of adrenalectomized rats to diminish. Parathyroidectomized rats with low blood calcium have a specific appetite for calcium. The mechanism of this process must be fairly complex; apparently, however, it does not involve a decrease in the absolute threshold of taste. Thus the minimum concentration of NaCl applied to the tongue evoking action potentials in the chorda tympani was found to be 0.008 per cent in normal and 0.010 per cent in adrenalectomized rats.

Fatigue soon occurs in taste. Persistence of a stimulus diminishes the sensitiveness of the taste stimulated. There is some evidence that sensi-

[1] FALCONER, D. S., *Ann. Eugenics*, **13**, 211, 1947.

[2] SNYDER, L. H., *Science*, **74**, 151, 1931; BLACKSLEE, A. F., *Proc. Nat. Acad. Sc.*, **18**, 120, 1932.

[3] RICHTER, C. F., *Endocrinology*, **26**, 367, 1939.

tiveness of other taste receptors is simultaneously increased.

Contrasts and masking. Simultaneous and successive contrasts are well developed in taste. Thus, sweet and acid reciprocally raise, and sweet and bitter lower, their thresholds. Neutral distilled water tastes like dilute alkali, but after an acid solution it tastes sweet. These contrasts are used to mask unpleasant tastes and to enhance pleasant ones. In lemonade, acid and sweet are combined in agreeable proportions. Aromatic substances are sometimes used to mask an unpalatable taste.

Fusion frequency. Repetitive electric stimulation of the tongue, with the cathode as the active electrode and the inactive anodic electrode on the lower surface of the tongue, gives rise to a sensation of sour taste. According to Allen and Weinberg,[1] if the frequency is sufficiently low, pulsating or discontinuous sensation is evoked, which becomes continuous when the frequency of stimulation is increased. The critical frequency, *i.e.*, the minimum necessary to evoke a continuous sensation, decreases as the voltage increases; a stronger stimulus prolongs the duration of the sensation. Plotting frequency against voltage, four curves are obtained which the authors attribute to the four tastes on the basis of experiments with gymnemic acid which abolishes the sweet and partly the bitter tastes. Jones and Jones,[2] however, have not been able to obtain discrete gustatory sensations; only continuous sensations were evoked with all voltages and frequencies. Tactile pulsating sensations were perceived which might be confused with gustatory sensations by untrained subjects. It is therefore premature to speak of fusion frequency in gustatory sensation.

Enzymes in taste organs. Histochemical methods have shown there are several enzyme systems in the taste buds or the neighboring epithelium of the papillae. Alkaline phosphatase has been found concentrated in all the cells of the taste buds, gustatory pores, and gustatory hairs in the bat and monkey; in man and rabbits it is found in the epithelium overlying the taste buds, gustatory pores, and gustatory hairs. Other enzyme systems have been found in the epithelium lining the folds between the papillae; they act on the following substrates: adenosinetriphosphate, hexose-diphosphate, adenylic acid, ribonucleic acid, long-chain fatty acids (lipase); acid phosphatase has also been identified. These enzymes have been

[1] ALLEN, F., and M. WEINBERG, *Quart. J. Exper. Physiol.*, **15**, 385, 1925.
[2] JONES, M. H., and F. W. JONES, *Science*, **115**, 355, 1952.

considered as part of the mechanism which stimulates taste receptors.[1]

Localization. Taste sensations are localized on the receptor. They are contact sensations similar in this respect to tactile sensations.

Peripheral and Central Pathways of Taste

Innervation of the receptors. The taste buds are innervated by fine fibers. In the cat 18 per cent, and in the dog 23 per cent, of the fibers in the chorda tympani (one of the gustatory nerves) are unmyelinated and less than 1.5 μ in diameter. Myelinated fibers range from 1.5 to 6 μ in diameter.[2] Not all are gustatory fibers; some of them conduct pain and tactile impulses. The following nerves include fibers which innervate taste receptors.

1. *The lingual nerve* innervates the anterior two-thirds of the tongue. The taste fibers in this nerve arise in the geniculate ganglion of the facial nerve. In most subjects they join the lingual nerve through the chorda tympani; in others they follow a different path, leaving the geniculate ganglion through the greater superficial petrosal nerve and passing through the otic ganglion into the chorda and lingual nerves.
2. *The glossopharyngeal nerve* innervates the posterior third of the tongue. The fibers arise in the petrosal ganglion.
3. *The vagus nerve* sends a few fibers to the taste buds of the epiglottis and pharynx. The neurons of these fibers are situated in the ganglion nodosum (Fig. 408).

Section of the lingual and glossopharyngeal nerves increases considerably the threshold for salt and quinine, but complete ageusia is not observed even though the taste buds of the tongue degenerate completely after denervation. Taste buds innervated by the vagus may be responsible for the remnant of sensation, or else the strong solutions employed stimulate nongustatory receptors (the common chemical sense).[3]

[1] BOURNE, G. H., *Nature, London*, **161**, 445, 1948; BARADI, A. F., and G. H. BOURNE, *Nature, London*, **167**, 977, 1951; *Science*, **113**, 660, 1951.
[2] FOLEY, J. D., *Proc. Soc. Exper. Biol. & Med.*, **60**, 262, 1945.
[3] PFAFFMANN, C., and J. K. BARE, *Am. Psychol.*, **3**, 284, 1948; PATTON, H. D., and T. C. RUCH, *Ann. Rev. Physiol.*, **12**, 469, 1950.

Central pathways and nuclei. The fibers conducting impulses from the taste buds, which enter the pons and medulla with the seventh, ninth, and tenth nerves, form part of the descending tractus solitarius and end in the nucleus of this tract. The facial and glossopharyngeal gustatory fibers end in the rostral and the vagal fibers in the caudal part of the nucleus. Lesions in the tractus solitarius (guinea pig, rabbit) produce degeneration in the dorsal part of the medial lemniscus, medial to the ventral trigeminothalamic tract. These fibers end in the n. ventralis posteromedialis, or arcuate nucleus of the thalamus. Lesions in the dorsal part of the medial lemniscus raise the threshold for quinine in the rat.

FIG. 408. Diagram representing the peripheral and central pathways of taste (highly schematic).

The thalamic relay neurons of the gustatory paths are situated in the dorsomedial tip of the arcuate nucleus. The following observations are evidence for this assertion: (*a*) bilateral localized lesions in the arcuate nucleus in the monkey are followed by permanent disturbances in taste; the threshold for quinine is raised, and the animals drink a bitter-tasting solution as if it were water;[1] (*b*) hemiageusia has been reported in a patient with a tumor destroying the medial tip of the arcuate nucleus;[1] (*c*) tactile stimulation of the body surface in the cat and monkey evokes potentials in the arcuate nucleus, except for a discrete area in the medial tip of this nucleus; (*d*) extensive lesions of the inferior rolandic cortex which do not cause disturbances in taste provoke severe retrograde degeneration in the arcuate nucleus except in its dorsomedial tip.

[1] RUCH, T. C., M. BLUM, and J. BROBECK, *Am. J. Physiol.*, 133, 433, 1941; PATTON, H. D., and T. C. RUCH, *J. Neurophysiol.*, 7, 171, 1944.

Cortical representation. It has been commonly supposed that the cortical centers of taste were situated in the olfactory area. This opinion was founded on the following facts: (*a*) taste and smell are "chemical" senses; (*b*) the receptors of both senses are frequently stimulated at the same time, and the sensations provoked by one and the other are closely associated; (*c*) stimulation of taste and smell produce motor and secretory reflexes in the digestive tract. In spite of this close functional association, the senses do not have a common or even neighboring cortical representation. Removal of olfactory centers (pyriform area, amygdaloid nucleus, and hippocampus) is not followed by disturbances in taste.[2] Extirpation of the olfactory bulb does not suppress gustatory reactions in animals, nor are the

[1] ADLER, A., *Ztschr. Neurol. & Psychiat.*, **149**, 208, 1934.
[2] ALLEN, W. F., *Am. J. Physiol.*, **132**, 1941, *loc. cit.*

taste thresholds modified when olfactory capacity is diminished, as occurs in the common cold and other cases of nasal obstruction.[1]

The central pathways and the thalamic and cortical representations of taste are not close to those of smell, but are associated with those of the somatic sensations of the tongue, the impulses of which are conducted in the trigeminal nerve.[2] The exact area of cortical representation of taste has not yet been definitely located, but there is the following evidence that it is not situated on the surface of the brain, but in the sylvian fissures in the parainsular region, or in the insula: (*a*) removal of the precentral and postcentral parainsular cortex with minimal damage to the surface cortex produces disturbances in taste in monkeys and chimpanzees as severe as those provoked by wide and deep lesions in the opercular region;[3] (*b*) stimulation of the chorda-lingual trunk, after section of the lingual component, evokes surface potentials in the cortex of the sylvian fissure (Woolsey's somatosensory area) rostral to the face area;[4] (*c*) in the rabbit, extirpation of Bremer's "masticatory" area, the cytoarchitectural homologue of the insula in primates, abolishes taste and provokes retrograde degeneration in the posteroventral thalamic nucleus (homologue of the primate arcuate nucleus);[5] (*d*) epileptic seizures with gustatory aura and disturbances in taste have been observed in patients with tumors involving the insula. Cortical representation of taste is, therefore, situated in the inferior extension of the sensory cortex, below the somatosensory representation of the face, corresponding to the extreme medial localization of the thalamic relay of the gustatory path.[6]

[1] Frequently it is said that food becomes tasteless when one suffers from a cold, because the olfactory sensations, which are diminished or absent, are mistakenly identified with those of taste, which are not impaired.

[2] Börnstein, W. S., *Yale J. Biol. & Med.*, **12**, 719; **13**, 133, 1941.

[3] Ruch, T. C., and H. D. Patton, *Federation Proc.*, **5**, 89, 1946.

[4] Patton, H. D., and V. E. Amassian, *Federation Proc.*, **9**, 99, 1950.

[5] Gerebtzoff, M. A., *Arch. internat. de physiol.*, **51**, 199, 1941.

[6] Patton, H. D., *Ann. Rev. Physiol.*, **12**, 469, 1950.

Physiologic Significance

Stimulation of taste receptors provokes salivary, gastric, and other digestive secretions and gastrointestinal motility. Taste thus plays an important, although not indispensable, part in normal digestion.

Taste, however, has a more fundamental physiologic significance; it regulates the ingestion of food and in this way has an important part in the maintenance of the constancy of the *milieu intérieur*. This fact has only recently been brought to light, mainly by the work of Richter.[1] Food is distinguished from harmful substances. Thus rats prefer nutritive solutions of sugars and electrolytes, below certain concentrations (8 per cent saccharose, 10 per cent maltose), and will drink them instead of distilled water; they also refuse toxic solutions, such as those of $HgCl_2$ and AsO_2, or substances without nutritive value. When given a choice between water and alcohol, they prefer alcohol in concentrations of 1 to 6 per cent but not in higher concentrations; they reduce the ingestion of food in quantities equivalent to the calories of the alcohol ingested. Rats on deficient diets show a marked preference for foods that contain the required substance. Thus rats suffering from vitamin deficiency will eat selectively foods containing the missing vitamin.

The intimate nature of this process is not yet known, but taste plays a fundamental part, because after section of the taste nerves preference thresholds and discriminative ingestion of food are considerably disturbed.

BIBLIOGRAPHY

Börnstein, W. S., Cortical Representation of Taste in Man and Monkeys, *Yale J. Biol. & Med.*, **12**, 719; **13**, 133, 1940.

Moncrieff, R. W., "The Chemical Senses," 2d ed., Leonard Hill, London, 1951.

Parker, G. A., "Smell, Taste and Allied Senses in Vertebrates," Lippincott, Philadelphia, 1922.

Patton, H. D., Smell and Taste, *Ann. Rev. Physiol.*, **12**, 469, 1950.

Richter, C. P., Physiological Psychology, *Ann. Rev. Physiol.*, **4**, 561, 1942.

[1] Richter, C. P., *Harvey Lect.*, **38**, 63, 1943.

CHAPTER 76

Vision

Knowledge of the size, shape, color, and luminosity of objects in the environment, together with their position and movements in space with respect to other objects and to the observer, is obtained by means of sight. Vision also plays an important part in postural reflexes and in the maintenance of postural equilibrium. These results are obtained by a complex process. Light from the objects in the environment is focused on the retina by the dioptric system of the eye, and an image is formed. The receptor cells in the retina are thus stimulated, and nerve impulses arise, which are conducted along the visual pathway to cortical centers. There the mental image of the object is formed. The process will be examined in this order in the course of this chapter and the following ones.

THE FORMATION OF IMAGES BY THE EYE

Light rays that enter the eye are focused on the retina by the dioptric surfaces in the eye according to the physical laws of optics.[1]

Refraction of light. Light rays, on passing obliquely from one transparent medium to another of different optical density, are deflected from their path, *i.e.*, they are refracted. According to the laws of refraction, when a ray of light passes from a rarer to a denser medium, refraction is made toward the perpendicular, and away from the perpendicular when the ray passes from a denser to a rarer medium. For example, the incident ray, on passing from air into a sheet of glass of greater optical density, is deflected toward the perpendicular; on passing out of the glass into air at the opposite surface, it is again deflected, but this time away from the perpendicular (Fig. 409). The incident ray forms an angle with the perpendicular to the surface which is called the *angle of incidence;* the refracted ray forms with the perpendicular the *angle of refraction*. The ratio of the sine of the angle of incidence to the sine of the angle of refraction is the *index of refraction* of the medium into which the ray enters.

[1] The student is advised to refresh his knowledge of optics in a textbook of physics.

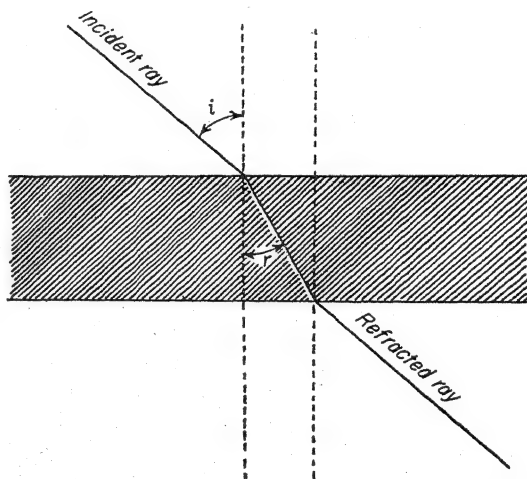



Fig. 409. Refraction of light passing from one medium to another of different density. *i*, angle of incidence; *r*, angle of refraction.

If a light ray falls obliquely on a transparent medium with parallel surfaces (*e.g.*, a sheet of glass), on again passing out of the medium into air it will follow a path that is parallel to the incident ray. If a ray falls obliquely on the surface of a prism, on again emerging into the air it will be deflected toward the base of the prism. If the ray passes through a lens, the deflection will depend on the form of the lens, *i.e.*, convex, concave, cylindrical, or complex.

Refraction by convex lenses. The principal axis of a convex lens with two spherical surfaces is a straight line that passes through the centers of curva-

ture of these spheres; therefore it is perpendicular to the principal plane of the lens. Rays that fall on the lens traveling along the principal axis are not refracted; they pass through the lens without being deflected. Rays parallel to the principal axis are refracted and on emerging converge to a point situated on the principal axis behind the lens; this is the *principal focus* (Fig. 410). The distance from the principal focus to the lens is the *principal focal distance;* it varies with the curvature and the refractive index of the lens. The focal distance is a measure of the refractive power of a lens.

If a luminous object is placed on the principal focus, the rays emerging from the lens follow a parallel course. If a luminous object is placed on the principal axis between the principal focus and the lens, the rays emerging from the lens follow a divergent course and are not focused. If a luminous object is placed on the principal axis at a greater distance than the principal focal distance, the rays on emerging from the lens converge to a point on the principal axis which is farther away from the lens than the principal focus. These two points are known as *conjugate foci,* because a luminous object placed on either of them will be focused on the other. For example, if a luminous object is placed at twice the principal focal distance, a real image, inverted and smaller than the object, will be formed at twice the principal focal distance behind the lens.

The *optical center* of a biconvex lens is situated on the principal axis equidistant from the surfaces if these are of equal curvature, and nearer one of the surfaces if they are of unequal curvature. All straight lines passing through the optical center, other than the principal axis, are called *secondary axes;* these cross the principal axis at the optical center. Rays of light coincident with a secondary axis do not suffer a change in direction; they emerge from the lens and follow a path that is parallel to the incident ray. A luminous point not on the principal axis will have its conjugate focus (*i.e.,* image) on the secondary axis drawn from this point through the optical center, at the place where the prolongation of the line of the refracted ray passing through the principal focus cuts the secondary axis.

Fig. 410. Refraction of light by convex lenses. *A*, refraction of parallel rays; *B*, refraction of divergent rays; *C*, refraction of divergent rays from a light placed between the lens and the principal focus; *f* and *f'*, conjugated foci; *a* and *a'*, inverse principal planes; *b* and *b'*, focal planes; *c*, direct principal plane.

The Refractive Surfaces of the Eye

The *cornea* and *aqueous humor* have approximately the same index of refraction; they can therefore be considered as forming a system of concavoconvex lenses. When this system is separated from the organism, its refractive index is lower than that of the crystalline lens, but in the intact eye it has a higher refractive power because light rays pass from the air into the much denser medium of the cornea. Two-thirds of the

total refractive power of the eye is due to this system (42 D.).[1]

Table 98. Refractive Properties of the Eye

	Refractive index	In the organism		Radius of curvature, mm.	Distance to cornea, mm.
		Refractive power, D.	Focal distance, mm.		
Cornea	1.34	42	24	8	
Aqueous humor	1.33				
Crystalline lens	1.42	23	44	Ant. surf. 10	3.6
				Post. surf. 6	7.6
Vitreous humor	1.33				
Retina	....	..	..	....	22.6

The *crystalline lens* is a biconvex lens which has a high refractive index but which *in situ* does not refract the light rays so much as when isolated from the eye, because the rays pass into it from the cornea and aqueous humor, which also have high optical density. After the crystalline lens has been removed because it has become opaque (cataract), a lens of 10 D. corrects the defect produced. The cortex of the lens has a lower refractive power than the nucleus, a circumstance that is useful for the correction of spherical and chromatic aberration produced especially in the periphery of the lens. The shape of the lens can be altered, thus modifying its refractive power. This can increase up to 10 D. when an object is displaced from a distance greater than 6 m. (20 ft.) from the cornea (far point) to a distance of 10 to 20 cm. (4 to 8 in.) from the cornea (near point). This accommodation of the lens, according to the distance at which the object is placed, permits clear focusing of the image on the retina. It is governed by a special mechanism, which will be considered later.

The reduced eye. The path of the light rays through the eye can be deduced from the refractive index and the surface curvature of the different media and the distance between them (Table 98). This is a laborious task, and a fair approximation can be obtained by employing an optical system, introduced by Listing and modified by Donders, known as the "reduced eye." The principal numbers in this system are easily remembered, and there is no difficulty in tracing the path of the light rays. Light is considered as being refracted only on entering the eye, which is supposed to be a homogeneous optical medium with the same index of refraction as water, 1.333. The center of curvature and optical center of the system is situated on the principal axis, 5 mm. behind the cornea; the retina is 15 mm. behind the optical center. The distance from the cornea to the retina is 20 mm., and parallel rays that enter the reduced eye are focused exactly on the retina. The total refractive power is 65 D. (Fig. 411).

[1] The refractive power of lenses is expressed in diopters, *i.e.*, the reciprocal of the focal distance in meters. Thus a lens with a focal distance of 1 m. is considered as having a refractive powerof 1 D., if the focal distance is 0.5 m. the refractive power is 2 D., and if the focal distance is 0.1 m. the refractive power is 10 D.

Retinal Image

A real, inverted image, smaller than the object, is formed on the retina. Scheiner observed this for the first time in 1625, using the eye of a white rabbit, which has no pigment and therefore has a translucid sclera. If the eye of any animal is placed at one end of a tube made out of black paper, with the cornea facing outward so that the light enters the eye, on looking through the tube at the other end, the observer will see the image of the objects before the eye formed on the retina. After scraping off the sclera, the result is even more striking. The retina can be examined *in vivo* by means of the ophthalmoscope.[1]

[1] The ophthalmoscope was invented by Helmholtz in 1851. The primitive model consists of a small mirror that reflects a ray of light into the eye. In the center of the mirror there is a small orifice through which the observer looks into the fundus of the eye.

Visual angle. Size of the retinal image. The path of the light rays forming the retinal image is given in Fig. 411. The visual angle is formed at the optical center by the limiting rays from *A* and *B*. This angle increases as the object is placed closer to the eye, *i.e.*, the visual angle is inversely proportional to the distance between the object and the eye. These same lines prolonged beyond the optical center form the retinal image. There are, therefore, two triangles joined at their apexes, the bases of which are respectively the object and its image. From the size of the object and its distance from the eye, it is easy to calculate the size of the image, because the distance of the image from the optical center is always 15 mm. Therefore,

$$\frac{\text{Size of object}}{\text{Size of image}} = \frac{\text{Distance from object to optical center}}{\text{Distance from image to optical center}}$$

For example, the image of an object of a diameter of 100 m. at 1,000 m. distance will measure 1.5 mm., since

$$\frac{100}{x} = \frac{1{,}000}{0.015}$$

$$x = \frac{0.015 \times 100}{1{,}000} = 0.0015 \text{ m. } (1.5 \text{ mm.})$$

Fig. 411. Formation of the retinal image in the reduced eye. Distances in millimeters. (*Donders.*)

Characteristics of the image perceived. Images are perceived "erect" and "projected." The retinal image is inverted with respect to the object, but it is seen in the correct position. The "righting" of the image is a psychological process, which is begun in childhood by the association of the visual image with other, especially with tactile, sensations.

Projection consists in the localization of objects at a distance from the receptor; touch and taste, unlike vision, are localized on the receptor. Thus if the eyes are closed and a finger presses on the eyeball, a visual sensation, called a "phosphene," is awakened; a dark spot surrounded by a light halo is perceived. The sensation is localized at a distance in a direction opposite to that of the stimulating pressure, *e.g.*, on the nasal side if the temporal side is stimulated.

Lecat's experiment is another illustration of the same process. A pinhole is made in a card, which is then placed at a distance of 2 cm. from the eye, *i.e.*, within the near point. The head of a pin is placed between the card and the eye, and they can be so adjusted that in the light passing through the pinhole an inverted pinhead will be seen. The pin casts an erect shadow on the retina, and an image is formed which corresponds to that of an inverted pin.

Entoptic images. Objects within the eye form what are known as entoptic images. Thus foreign bodies in the vitreous humor are seen as flying shadows when one looks at the sky or on a luminous surface. Defects on the cornea or crystalline lens appear as colored halos. The vascular network of the retina can also be perceived by looking at a bright luminous surface through a pinhole in a card which is moved so that the shadows of vessels on the retina are displaced (Purkinje's images). The erythrocytes in the retinal vessels can be seen by looking through a blue-violet glass onto a bright surface, or even better with Fortin's apparatus, which consists of a bright blue-violet light placed laterally to the eye and moved backward and forward.

Optical Defects

The eye has defects common to all optical systems. These are, however, corrected to a certain extent by the iris, which eliminates marginal rays (the cause of the majority of these defects), and by the low sensitiveness of the peripheral retina, where the distorted images are mostly formed. There are also psychological processes

which contribute to improve the appearance of the images.

Spherical and chromatic aberration. Diffraction. Presbyopia. *Spherical aberration* in a lens is due to focusing of marginal rays in front of the focus of the central rays, thus blurring the image. In a camera this defect is corrected by forming the lens with two substances of different refractive indexes. In the eye it is corrected in part by elimination of the peripheral rays (constriction of the iris) and by the greater optical density of the nucleus of the lens with respect to that of its cortex.

Chromatic aberration is due to the different dispersion of the light rays by the lens, according to their wavelength, so that the focus for violet rays is placed nearer the lens than that of red rays, with a series of intermediate foci for the other colors of the spectrum. This type of aberration is corrected by joining a concave lens to a convex one of a different type of glass, so that the composite lens is achromatic. In the eye yellow rays are focused on the retina so that 70 per cent of the spectrum is utilized. Colors at both ends of the spectrum are not perceived separately, nor is the image blurred. The mechanism of this is not well understood.

Diffraction consists in the deviation of part of the rays by the edge of the iris or by colloidal particles in the cornea or the crystalline lens. This blurs the retinal image. The ultraviolet rays are those which suffer the most dispersion; thus they are deviated from the retina, on which they have a harmful effect.

Presbyopia consists in the decrease of the power of accommodation of the lens for objects at different distances. This power diminishes fairly rapidly between forty and fifty years of age, and the near point of distinct vision is situated at increasingly greater distances from the eye. It is easily corrected by lenses that replace the diopters lost.

Emmetropia and ametropia. In the normal eye, when the lens is at rest, parallel rays are focused on the retina. This is due to the existence of an adequate relation between refractive power and axial length, which is called emmetropia. Normally the far point is situated at 6 m. (20 ft.), and rays of light from that or a greater distance can be considered as coming from infinity and entering the eye parallel to the principal axis.

If there is an abnormal relation between refractive power and the axial length of the eye, the condition is known as ametropia. The far point is not placed at infinity, and parallel rays are not focused on the retina. Ametropia may be due to the fact that the axis is too long, in which case it is called *myopia*, or too short, in which case it is *hyperopia* or *hypermetropia* (Fig. 412).

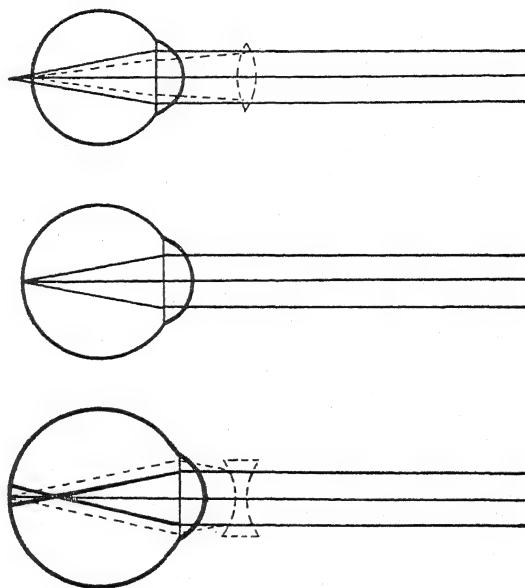



FIG. 412. Hyperopia, emmetropia, and myopia. Course of light rays and formation of retinal image in the normal eye (center), hyperopia (above), and myopia (below). Mechanism of correction by appropriate lenses.

Myopia. In "short-sighted" people the axial length of the eye is too long with respect to the refractive power of the eye. Parallel rays are focused on some part of the vitreous humor, and a blurred image is formed on the retina. Only divergent rays from an object placed near the eye are focused on the retina. The far point and the near point are close to each other, and accommodation does not play an important part; therefore the ciliary muscle tends to atrophy. The defect is corrected by the use of concave lenses.

Hyperopia. In this case the eye is too short for the refractive power of the eye, and parallel rays are focused behind the retina. The retinal image is blurred, and the individual must accommodate even to focus distant objects. This, together with the effort of accommodation required for near objects, leads to hypertrophy of the ciliary muscle and eventually to eyestrain. The defect is corrected by convergent lenses.

At birth the majority of eyes are hyperopic. Later the eye increases in length more rapidly than it decreases in refractive power, and it tends to become less hyperopic or even myopic. Hyperopia is therefore gradually corrected with age, while myopia increases.

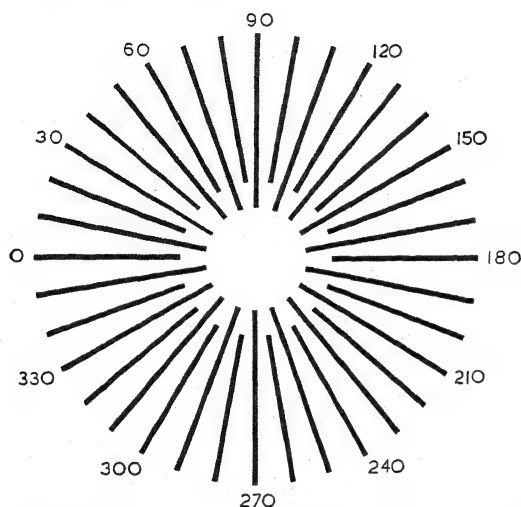



Fig. 413. Astigmatic chart of Lancaster-Regan. The normal eye sees all the meridians clearly. In astigmatism some of the meridians appear blurred.

Astigmatism. In a spherical lens all parallel rays converge to a single focus. In toric lenses, *i.e.*, those with one meridian of smaller curvature than that of the meridian placed at right angles to the former (similar in shape to the outer surface of an automobile tire), a single focus is not formed. The rays falling on the meridian with greatest curvature converge to a point placed in front of those falling on the meridian of smallest curvature. The image is therefore distorted. The cornea of certain individuals is not perfectly spherical but to a certain extent toric in shape. These individuals see clearly lines placed on one of the meridians, but not those on others (Fig. 413). This defect is known as regular astigmatism. It is corrected by cylindrical lenses placed at an adequate angle. If there is considerable irregularity of the surface of the cornea (irregular astigmatism), it is sometimes impossible to obtain a satisfactory correction.

VISUAL SENSIBILITY

The sensibility of the eye as a receptor of light rays is so great that it approximates the maximum theoretically possible.

The Receptor. Dark and Light Adaptation

The eye undergoes certain gradual changes when it passes from light to darkness or vice versa, which give it peculiar properties adapted for vision in either circumstance.

Dark adaptation. Scotopic vision. When an individual passes from a well-lighted environment to a dark one, at first nothing is seen, but after a few minutes there is a certain degree of vision, *i.e.*, the outline of objects is distinguished. The improvement in vision is due to a considerable increase in the sensitiveness of the retina, especially of its peripheral extrafoveal part. The minimum amount of light that can be perceived by a fully dark-adapted retina is 1:10,000,000,000 of the maximum stimulus in full daylight.

Curve of dark adaptation. In the course of dark adaptation the pupil dilates, the visual purple increases, and the chromatosomes are concentrated. The green part of the spectrum appears more luminous, with a maximum of 5150 A, while in a bright light maximum luminosity corresponds to the yellow part of the spectrum (6100 A).

If the threshold for light is measured at intervals after passing from bright light into darkness, it will be seen to fall sharply during the first 5 min.; then, after a short interval during which the curve tends to flatten, it again falls gradually, becoming stable after 20 to 30 min. (Fig. 414). The first drop is due to an increase in the sensitiveness of the cones, the threshold of which remains stable at this level. The second, gradual fall, which is more marked, is due to an increase in the sensitiveness of the

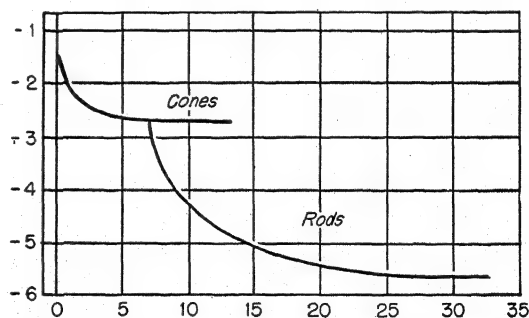



Fig. 414. Curve of dark adaptation. Ordinate, logarithm of threshold intensity in millilamberts. Abscissa, time in minutes.

rods. Nocturnal vision is due mainly to stimulation of the rods, and maximum sensitiveness corresponds to the peripheral part of the retina, not to the fovea. For this reason objects are perceived better in a slightly lateral position than when they are looked at directly.

Factors influencing dark adaptation. A dark-adapted eye loses its adaptation if a bright light falls on the eye even for only a brief instant. raising the hands to cover the eyes. A series of changes opposite to those of dark adaptation takes place (Table 99), and after a few minutes the eyes become light-adapted. The process is much shorter than that of dark adaptation.

The double function of the retina. The retina functions differently in daylight and at night. Schultz in 1866 was the first to observe this.

Table 99. Adaptation of the Eye to Light and Darkness

Adaptation	*Pupil*	*Retina*				*Sensitiveness*				*Avitaminosis A*
		Visual purple	*Melanosomes*	*Part used*	*Receptors*	*To light*	*Visual acuity*	*To luminosity of spectrum (Purkinje phenomenon)*	*To radiant energy (apex of visibility curve)*	
Light-adapted......	Myosis (3 mm.)	Decreased	Dispersed	Fovea	Cones	Diminished	Great	Yellow	554 mμ	Normal vision
Dark-adapted......	Mydriasis (8–9 mm.)	Increased	Concentrated	Peripheral retina	Rods	Increased	Small	Green	510 mμ	Night blindness

Aviators who have to fly at night are advised to wear red goggles for some time before the flight, because these permit central vision by means of the cones without interfering with the process of dark adaptation. Light has a considerable effect on this process, which is greatly delayed if the eye has been previously submitted to intense or prolonged illumination.

Anoxia raises the threshold for light. Thus an ascent of 5,000 m. (16,500 ft.) increases the threshold two and a half times. On the contrary, hyperventilation at sea level causes a drop in the threshold to half the normal value.

Vitamin A deficiency causes great difficulty in nocturnal vision. The whole curve of dark adaptation is raised. This is due to a disturbance in the formation of visual purple, which is cured by the administration of vitamin A if the deficiency has not been excessively prolonged.

In a rare condition, often hereditary, called "nyctalopia," there is also a disturbance in dark adaptation and night blindness. It is not improved by vitamin A.

Light adaptation. Photopic vision. On passing from a dark or poorly lighted environment to a bright light, there is a disagreeable sensation of dazzling, which may be painful and provokes defensive reflexes, such as blinking and

Daylight vision, which gives color and detail, is performed by the cones, mainly in the fovea, and has special nerve paths. Nocturnal vision gives only a colorless outline of the objects, but a minimum amount of light energy can stimulate the receptors, *i.e.*, the rods, especially in the peripheral part of the retina. Visual purple plays an important part in this type of vision, which also has special nerve paths. Other differences between daylight and nocturnal vision are given in Table 99. In the eye of higher animals, therefore, there are two different, but correlated, visual processes.

The Stimulus. Intensity, Functions

Light is the specific stimulus for the visual receptor, but the eye is sensitive only to rays of certain wavelengths (as will be seen when considering color vision), and within those wavelengths the stimulus must have a certain intensity. The intensity of the visual stimulus, brightness, which can be accurately measured by physical methods, should be distinguished from the intensity of the visual sensation, luminosity, which is a complex psychological fact. The intensity of the latter is related to that of the former, but is not exclusively dependent on it.

Three thresholds must be considered: (*a*) the absolute threshold; (*b*) the difference threshold; (*c*) the critical fusion frequency.

The absolute threshold is the minimal amount of light that can be perceived. It is dependent on several factors, among which are the wave-

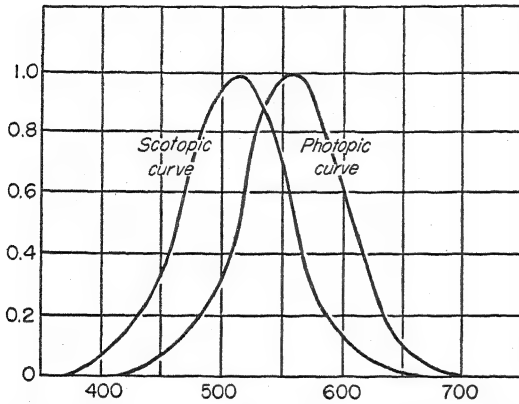



Fig. 415. Luminosity curves of spectra for the eye adapted to light (photopic) and darkness (scotopic). Ordinate, coefficient of visibility. Abscissa, wavelength in millimicrons.

length of the rays and the state of adaptation of the eye. This is well demonstrated by the determination of visibility or luminosity curves for different wavelengths, using a constant amount of energy (Fig. 415). These curves show that the eye is most sensitive to the middle part of the spectrum, although the maximal point is different in daylight and nocturnal vision. The curves in the figure have been adjusted so that the maximum in each of them corresponds to a conventional value of 1. The minimal threshold for nocturnal vision is far below that for daylight vision. The sensitiveness of a dark-adapted eye is the maximum theoretically possible. The duration of the stimulus, which should not be less than ⅛ sec., and the size of the area of retina stimulated modify the absolute threshold. Several laws (Ricco's law, etc.) define the relations between the duration of the exposure, the size of the area stimulated, and the absolute threshold.

The difference threshold is the minimal difference of intensity that can be perceived. The same factors that modify the absolute threshold have an influence on the difference threshold.

The critical fusion frequency is the minimum frequency at which intermittent flashes of light produce a continuous sensation. For example, if a disk, out of which a sector has been cut, is placed before a source of light and rotated at a slow rate, intermittent flashes of light will be perceived; if the rate of rotation is increased, a continuous brightness will be experienced. The minimum speed of rotation, *i.e.*, the minimum frequency at which the flashes become fused into a continuous sensation, is the fusion frequency. Fusion frequency, therefore, measures the maximum number of stimuli that can be perceived separately in unit time. Several factors condition fusion frequency; it is constant for a given intensity but rises with the increase in intensity (Fig. 416). Fusion frequency is not dependent on color; therefore it can be used to compare the brightness of lights of different color. Critical fusion frequency diminishes with age, fatigue, hypoxia, certain drugs (*e.g.*, alcohol), etc.

The curve obtained by plotting critical fusion frequency against intensity (Fig. 416) is smooth when the light falls on the fovea. When it falls on other parts of the retina, the curve has two limbs; the lower limb is interpreted as corresponding to the response of the rods and the upper limb to the response of the cones.

Visual Acuity

Visual acuity is the preciseness with which the detailed form of objects is perceived. It is a

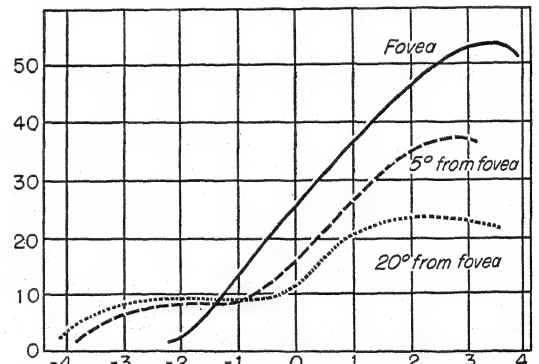



Fig. 416. Curves of critical fusion frequency for the fovea and at 5° and 20° from the fovea. Abscissa, logarithm of light intensity in photons; ordinate, fusion frequencies. (*Hecht and Verrijp.*)

function of daylight vision and is closely dependent on good illumination.

There are two ways of measuring visual acuity: (*a*) the determination of the *minimum visible;* (*b*) the determination of the *minimum*

separable. The minimum visible is expressed by the narrowest line that can be distinguished on a homogeneous background. The minimum separable is the smallest distance that must separate two parallel lines so that they can be clearly distinguished from each other. It is expressed by the *visual angle, i.e.*, the angle formed by the rays converging from the lines to the optical center. Most normal subjects have a visual angle of 1′, but in certain individuals with greater visual acuity, it measures less. The visual angle increases as visual acuity diminishes.

Factors modifying the visual angle. Three groups of factors modify the visual angle: (*a*) factors dependent on the stimulus; (*b*) dioptric factors; (*c*) retinal factors.

Factors dependent on the stimulus. Illumination is a most important condition for visual acuity. Visibility increases when the brightness of the object contrasts with the brightness of the background. It also increases with the intensity of illumination, but only up to an optimum, because excessive illumination produces glare. These factors can be demonstrated by varying the size, contrast, and illumination of black letters drawn on a white background.

Dioptric factors. The normal retinal image is not perfect; it is distorted by spherical and chromatic aberration and by diffraction by the optic media of the eye. In abnormal conditions, errors of refraction are added (myopia, hyperopia, astigmatism). Constriction of the pupil by excluding the marginal rays diminishes these errors and increases visual acuity. Monochromatic light improves visual acuity by diminishing chromatic aberration. A yellow light is the most favorable, then green, red, and blue light in decreasing order.

Retinal factors. The light-adapted eye has very great visual acuity for rays falling on the fovea. The acuity diminishes to one-half on the margin of the macula lutea and to one-fortieth in the extrafoveal parts of the retina. The dark-adapted eye has no visual acuity in the fovea and a low visual acuity in the extrafoveal parts of the retina (Fig. 417).

Visual acuity is mainly a function of the cones. It is at a maximum in the fovea, where the cones are closely packed, and in daylight, when the cones are functioning. Nocturnal visual acuity is probably due to the rods. It does not exist in the fovea, where there are no rods, and is found only in the peripheral retina.

Retinal-grain theory. This theory, proposed by Hecht, endeavors to explain the sensibility of the retina. The sensitiveness of a photographic plate depends on the size of the grains on its surface. By analogy, sensitiveness of the retina would be dependent on the size of its "grains," *i.e.*, the cones and

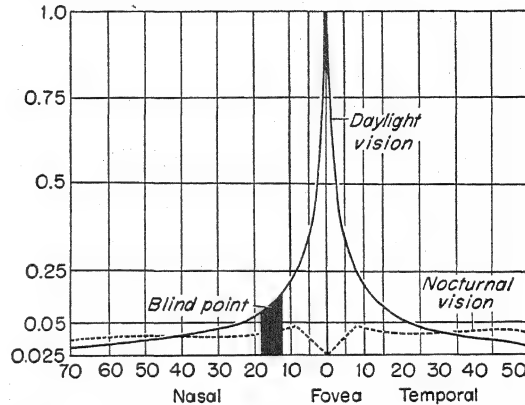



FIG. 417. Relative visual acuity curve in the central and peripheral retina.

rods. According to this theory, visual acuity depends on the difference in the illumination of neighboring rows of cones (granules), which would then be analyzed by the nerve centers. Thus, to perceive the minimum separable of parallel lines, two rows of cones would have to be in the shadow cast by these lines, and the cones between them would have to be brightly illuminated. In the case of the minimum visible, the thread or line would have to throw a shadow on a row of cones while the neighboring rows on both sides would be brightly illuminated. Visual acuity would therefore be due to a process of discrimination between light and shadow falling on the cones.

Careful measurements made by Polyak in 1941, on the distance between the cones in different parts of the retina, have shown that these distances are in agreement with the respective visual acuity only in the fovea. In other parts of the retina, visual acuity is not dependent on a single cone or rod but on the group of photoreceptors innervated by a single nerve fiber, *i.e.*, a sensory unit. The retinal-grain theory therefore should be based on functional, not on anatomic, units or "granules."

Clinical tests of visual acuity. Snellen's chart is commonly used to measure visual acuity in man. This chart consists of lines of letters of decreasing size, so that when read at distances of 60, 24, 18, 12, 9, 6, 5, and 4 m. the visual

angle measures 1′. The width of the lines forming each letter is one-fifth the total size of the letter.

The subject tested is placed at a distance of 6 m. (20 ft.) from the chart and indicates the smallest type of letter he can read. Visual acuity

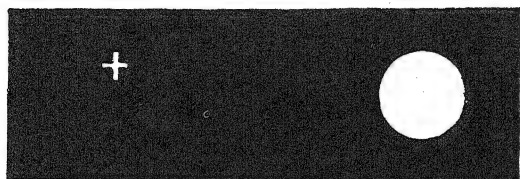



FIG. 418. Blind spot in the retina. The left eye is shut and the cross is looked at with the right eye. The figure is slowly moved away from and toward the eye. At a distance between 15 and 20 cm., the white disk is not seen

is expressed by the formula $V = d/D'$, in which V is the visual acuity, d is the distance of 6 m. and D' is the distance in meters corresponding to the line and noted on one side of the chart. For example, if the smallest type that can be read at 6 m. is that corresponding to 12 m., visual acuity will be equal to $6/12 = 1/2$. If the smallest line is that corresponding to 24 m., visual acuity will be $6/24 = 1/4$.

VISUAL FIELDS AND BINOCULAR VISION

The visual field of each eye is the extent of the environment that can be seen by a subject without moving the eye. It extends approximately 160° in the horizontal plane and 145° in the vertical plane. It is limited by the prominence of the nose on the mid-line or nasal plane, by the brow of the eye in an upward direction, and to a lesser extent by the cheek in a downward direction.

Objects within the visual field form an image on the retina. Those on the right of the subject are projected upon the left side of the retina, those above upon the lower half of the retina, and so on; *i.e.*, the visual field is inverted upon the retina. Therefore a subject with a blind right half of the retina loses the left half of the visual field.

The point in the retina at which the optic nerve enters the eye has no cones or rods; it is a blind spot. This spot is situated 3 mm. from the mid-line and 0.5 mm. below the fovea centralis (15° toward the temporal side of the fovea in the visual field). It is easy to demonstrate by a simple experiment that there is no vision when the image of an object is formed on this spot (Fig. 418).

Different parts of the surface of the retina are not equally sensitive to color (see "Color vision," below).

The visual field can be charted by means of the instrument known as the perimeter. The chin of the subject is placed on a support, and the eye is fixed on a point in front of it. A white or colored disk is moved along an arc, which can be fixed in any of the meridians of the eye field. The disk is moved along the meridian, and the subject indicates the moment when it enters the eye field. The results obtained from different meridians are marked on a chart, and the whole field can thus be plotted (Fig. 419).

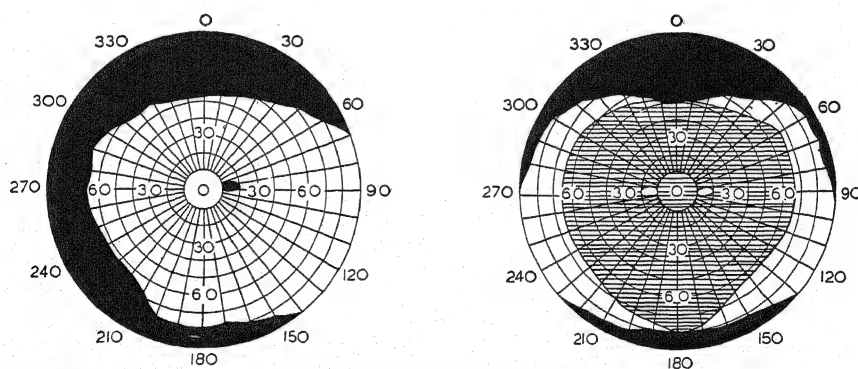



FIG. 419. Normal visual fields. On the left, visual field of right eye. On the right, binocular visual field.

Binocular vision. The visual field of one eye overlaps that of the other over an area of approximately 120°. Outside the area common to both eyes, each eye has a field, not visible to the other eye, which has the shape of a crescent con-

cave toward the nasal plane (monocular crescent or temporal half-moon) (Fig. 419).

Objects placed in the area common to the fields of both eyes form images on both retinas. They produce a single or a double visual sensation according to the position in the retinas on which the images are formed. The images produce a single visual sensation when they are formed on corresponding points of the retinas. Symmetrically placed points on the retina are corresponding points. Thus the half of the retina on the right of the fovea of one eye corresponds to the right half of the other; the two left halves, the two upper halves, and the two lower halves are corresponding. The movements of the eye are made so that the images of objects will fall on corresponding points of both foveas.

When the images are formed on noncorresponding points a double visual sensation, called "diplopia," is produced. Diplopia can be easily provoked by exerting light pressure with a finger placed on one of the eyeballs so that its axis of fixation is displaced. Objects within the field of vision will be seen double. In cases of strabismus (squint) the images fall on noncorresponding points, but there is no diplopia because one of the images is suppressed or ignored.

Binocular rivalry. When a different image falls on each eye, the images are not fused; one prevails over the other and is the only one perceived. Sometimes the images are seen alternately, or parts of the images prevail over the reciprocal parts of the others. This phenomenon is known as binocular rivalry. It does not obtain in certain cases of color vision in which fusion of the images gives rise to a different color (see "Color vision," below).

Stereoscopic vision. The images on corresponding points of the retinas are not absolutely identical. This slight difference is one of the elements used by the cortex to build up the notion of depth or solidity (*i.e.*, the third dimension) of the objects seen. Other data are also used for this purpose, *e.g.*, the size of the shadows or parallax, *i.e.*, the apparent displacement of near objects in the opposite direction to the movements of the observer, and of distant objects in the same direction. The notion of depth can be given by monocular, but is far more perfect in binocular, vision.

FIG. 420. Optical illusions. 1, the different segments of the vertical lines are all of equal length, although some appear to be shorter than others; 2, both figures are perfect squares of the same size, although they appear to be rectangles; 3, the distance from *A* to *B* is equal to that from *B* to *C*; 4, the white and black disks have the same diameter, but the white disk appears to be larger than the black one (irradiation); 5, the long horizontal lines are straight parallel lines.

PERCEPTION OF MOVEMENT

Perception of a moving object is maximal in light and when the image falls at a distance 10 to 15° from the fovea. It is a well-known fact

that the lateral vision of an object gives a better idea of its movements than central or foveal vision. In this case the smallest movement perceived is a displacement at an angle of 1 in 1 sec. if there are stationary objects in the neighborhood. If there are no other objects, the movements must be 10 to 20 times greater to be perceived.

If the eyes are fixed and the head or the whole body is moved, objects in the visual fields seem to move, owing to the displacement of their images on the retina. Thus when the observer is traveling in a train, stationary objects near the track seem to move in the opposite direction, and objects farther away in the same direction, as the train.

Optical Illusions

Errors of appreciation are committed with respect to the size, shape, and distance of objects if they are presented in certain ways. Figure 420 gives several examples of optical illusions.

The error known as "irradiation" is of special importance. It consists in the apparent increase in size of the image of a brightly illuminated object when placed next to one of the same size that is poorly illuminated. Thus in Fig. 420 the white circle seems to be larger than the black circle, although their diameters are equal. This effect is due to diffusion of stimulation around the edges of the image of the brightly lit object, owing to chromatic and spherical aberration.

COLOR VISION

Color vision is one of the most interesting aspects of the sensibility of the photoreceptor. It is a function of the light-adapted eye and is dependent on the acuity of the cones. Certain physical properties of light must first be examined in order to facilitate the understanding of the physiology of color vision.

Properties of Light as a Visual Stimulus

Light consists in radiations which on stimulating the photoreceptor give rise to sensations of color, which vary according to the intensity of the stimulus, the wavelength of the different radiations, and the combination of radiations of different wavelength.

The chromatic series. All visible colors are found in the solar spectrum with the exception of purple, which is the only extraspectral color. Colors are disposed in the spectrum according to their wavelength, forming the chromatic series. Every color has three physical attributes, *i.e.*, tone or hue, brightness or luminosity, and saturation[1] or purity.

Tone or hue. This attribute is due to the spectral distribution (*i.e.*, the wavelength) of the stimulus. The tones or colors have the following limiting wavelengths (in mμ): red, 723 to 647; orange, 647 to 585; yellow, 585 to 575; green, 575 to 492; blue, 492 to 455; indigo, 455 to 424; violet, 424 to 397. Light from a limited portion of the spectrum is known as monochromatic light.

Brightness or luminosity. A tone or hue varies in brightness according to the intensity of radiation. Excessive brightness causes an unpleasant sensation (glare). Brightness is measured by determining the critical fusion frequency, considered above. In daylight vision, yellow is the brightest color.

Saturation or purity. This attribute is dependent on the amount of accompanying white sensation. The purity or saturation of a color increases as the amount of white diminishes. Thus red is more saturated than pink. Probably in the retina a saturated color unsaturates other colors.

The degree of saturation is determined by means of the spectroscope. A saturated color gives out radiations within narrow limits of wavelength. If the color is not saturated, *i.e.*, if it is mixed with white, the dominant radiation is accompanied by radiations of other wavelengths. Thus pure red gives out radiations corresponding to the red part of the spectrum only, while unsaturated red gives out these radiations and those corresponding to other colors.

Color mixture or fusion. Color mixture refers here to the combination or fusion of spectral color, *i.e.*, of light radiations of different wavelengths, not to mixtures of pigments or paints

[1] The Optical Society of America has given the following definitions of these characteristics: Tone is the attribute of certain colors by which they differ in a characteristic way from gray of equal brightness and which classifies them as reddish, yellowish, greenish, or bluish. Brightness is the attribute of all colors by which a color may be classified as the equivalent of a given member of the series of grays going from black to white. Saturation is the attribute of all colors with a hue, which determines the degree by which they differ from gray of equal brightness (*J. Optic. Soc. America*, **6**, 527, 1922).

that may react chemically with each other and give rise to substances with new colors. In optics colors are mixed by means of lenses or mirrors that converge two or more light rays of different color (*i.e.*, wavelength) on a reflecting surface. When radiations corresponding to different colors fall on the same point of the retina, they produce a sensation that differs from the sensations produced by each radiation acting separately.

Primary colors. If a ray of sunlight is dispersed by a prism, the spectral colors can be recombined, and white light is again obtained without any loss of brightness (Abney's law). White light and any of the intermediate colors of the spectrum can also be obtained by fusion of red, green, and blue in certain proportions; these three are therefore called primary colors.

Complementary colors. Certain pairs of colors combined in adequate proportions give rise to a sensation of white; they are known as complementary colors. There is a complement for every color of the spectrum, *i.e.*, light radiations of a certain wavelength which, combined with the spectral color, give white light. Thus the following pairs are complementary colors: red (656 mμ) and greenish blue (492 mμ); yellow (595 mμ) and indigo blue (433 mμ); greenish yellow (563 mμ) and violet (less than 433 mμ). The complementary color for green is extraspectral purple, formed by the fusion of red and violet.

Fusion of pairs of noncomplementary colors. Complementary colors are separated by a fairly wide interval in the spectrum. Fusion of colors separated by a shorter interval gives rise to an intermediate color; *e.g.*, red and yellow give orange, red and blue give yellow.

The site of color fusion. Fusion of light rays of different colors apparently takes place in the brain, not in the eye. Thus if a ray of one color is made to fall on a certain spot of the retina of one eye, and a ray of another color on the corresponding point of the retina of the other eye, the colors are fused; this is known as binocular fusion. It is not yet clear which nerve centers are responsible for fusion; probably the geniculate bodies or the occipital cortex play a part in this phenomenon.

Achromatic series. White, black, and intermediate grays formed by mixtures of different proportions of white and black constitute the achromatic series. White is always obtained by fusion of radiations of different wavelengths: (*a*) by the resynthesis of the whole spectrum; (*b*) by fusion of the three primary colors; (*c*) by fusion of a pair of complementary colors. The origin of the sensation of black is not quite so clear. Some observers consider that black is not a sensation, because a black object does not reflect light. The sensation of black would therefore be due to the absence of stimulation of the retina. The image of a black object must, however, fall on some part of the retina to give rise to the sensation of black; if it falls on the blind spot, not black but nothing is seen. Moreover, black can be mixed with other colors and changes them. When the light stimulus ceases, the retina sends out electrical discharges; blackness, therefore, seems to be due to activity provoked by the cessation of the light stimulus.

Characteristics of the Receptor

The eye as a receptor. Each one of the primary colors is perceived from a definite part of the retina. If the visual fields for the different colors are determined by means of a perimeter, four concentric ringlike areas can be differentiated around the fovea. The innermost area corresponds to green, and the outer areas to red, blue, and white in that order, the white area extending to the ora serrata. The green area is trichromatic; all colors are perceived within this area. The red area is dichromatic; green is not perceived within it. The blue area perceives only blue (except for a narrow inner ring, which also perceives yellow).

The component rays of white light are refracted differently according to their wavelength. In the eye yellow light is focused exactly on the retina; rays of shorter wavelength are focused in front, and those of longer wavelength behind. This chromatic aberration is corrected in part by constriction of the pupil, thus eliminating the more peripheral rays, which undergo a greater degree of deviation than central rays.

Sensibility of the eye. The following aspects of the sensibility of the eye will now be considered: (*a*) spectral limits of visibility; (*b*) intensity threshold.

Spectral limits of visibility. The eye perceives all the colors in the solar spectrum between red (720 mμ) and violet (350 mμ). It is therefore sensitive to radiations of wavelengths differing by approximately one octave. The ear can

perceive sound waves within a range of approximately 10 octaves.

Beyond the red end of the spectrum, there are the infrared rays, which are not visible, *i.e.*, do not act on the retinal pigments but can be demonstrated by their thermic effect (*e.g.*, by means of a thermocouple). Beyond the violet rays there are ultraviolet radiations, which are also invisible but produce photochemical effects (*e.g.*, on a photographic plate). Rays with a wavelength between 350 and 320 mμ cause fluorescence in the eye media and thus become visible. With age the lens yellows, causing the violet end of the spectrum to become progressively invisible. If the lens is removed, violet is again perceived.

FIG. 421. The Purkinje phenomenon. Luminosity curves for spectra of different intensity. Abscissa, wavelength in millimicrons; ordinate, luminosity. (*König.*)

The difference threshold for color, *i.e.*, the capacity to discriminate color tones, varies for the different colors. Small differences in wavelength are perceived in the yellow part of the spectrum; the difference must be much greater in the red or violet parts to be perceptible. The normal human eye distinguishes approximately 165 tones or hues.

Intensity threshold. The Purkinje phenomenon. If luminosity is gradually diminished, the colors of the spectrum do not disappear all together, but one after the other. There is a chromatic threshold. Red and violet are the first to disappear and yellow, green, and blue the last. The threshold is therefore higher for the former colors than for the latter. The threshold for each color is constant. If the intensity is gradually increased above this threshold, 660 degrees of brightness can be distinguished for each hue.

If the brightness of the chromatic spectrum is diminished progressively, the colors gradually fade out, beginning at the ends of the spectrum; only a gray luminosity persists, more marked in the green part of the spectrum. This is known as the scotopic spectrum (dark spectrum), while the chromatic spectrum is also called the photopic spectrum (light spectrum). The brightest part of the photopic spectrum corresponds to the yellow wavelengths, while that of the scotopic spectrum corresponds to the green. This displacement of maximum brightness from the yellow (610 mμ) to the green (535 mμ) wavelengths, as illumination decreases, is called the Purkinje phenomenon (Fig. 421). This explains why yellow flowers appear the brightest at midday and blue flowers in the evening.

After-images. This term is applied to images that remain after the stimulus has ceased to act on the retina. They are especially notable after looking at a bright object. The after-image at first increases in intensity, then gradually disappears (Fig. 422).

FIG. 422. Diagrammatic curve of visual sensation following stimulation by a flash of light. (*After Hartridge.*)

There are two kinds of after-image according

to the conditions that obtain after the primary stimulus has ceased to act. If, after looking at a brightly illuminated object, the eyes are closed or fixed on a black surface, the after-image appears to be bright and of the same color as the object. This is a positive after-image, due apparently to a persistent effect of the stimulus. If, after looking at the object, the eyes are fixed on a white surface, the after-image takes on the complementary color. For example, if a bright red object is looked at for 20 sec. or more, the after-image will appear bluish green. This is a negative after-image, due to fatigue which causes an increase in the threshold for light of the wavelength of the stimulus and lowers the threshold for the complementary wavelength in the white light reflected by the surface. The retina therefore persists in an excited condition for a short time after the stimulus has ceased (positive after-image); the excited area is, however, less sensitive to the same stimulus and more sensitive to the complementary stimulus.

Contrast phenomena. A ray of light not only stimulates the retina; it also modifies the sensibility of the stimulated and neighboring areas.

Simultaneous contrast (*spatial induction*). Stimulation of a definite area of the retina causes an increase in sensitiveness to certain other stimuli in the neighboring areas. The maximum effect is produced with respect to the complementary color. Thus if a blue disk is placed on a yellow background both colors appear to be heightened by the contrast; if they are seen separated by a wide space they appear to be paler, *i.e.*, less saturated. The influence of color contrast is of importance in painting and in the employment of colors for decoration.

Successive contrast (*temporal induction*). Stimulation of the retina modifies its sensibility for successive stimuli. Thus after looking at a bright red disk, a green disk will appear to be more saturated. This is due to an increase in sensitiveness to the complementary color, as was mentioned in the discussion of after-images.

The Young-Helmholtz Theory of Color Vision

Many theories have been proposed to explain the phenomena of color vision, but none gives a completely satisfactory explanation of all the facts. In 1801 Thomas Young proposed a theory which was later expanded and modified by Helmholtz. It is known as the Young-Helmholtz theory and has wide acceptance.

According to this theory there are three different chemical substances in the cones, which are sensitive to red, green, and violet light respectively. Each one of these substances is specifically disintegrated by light of the corresponding color; other colors have a less considerable effect. The photochemical response stimulates the endings of different nerve fibers and sends impulses along the optic nerve to different parts of the occipital cortex according to the nature of the receptor that has been excited.

The sensations of red, green, or violet are due to the specific effect of the colored light on the corresponding chemical substance. The sensations of white and other colors of the spectrum are produced by the combined decomposition of the three substances in varying proportions. Black is due to the absence of stimulation.

This theory gives a good explanation for many facts. Thus, blindness to one color would be due to the absence of one of the photochemical substances. Color mixture would be produced by simultaneous decomposition of two of the three substances. The positive after-image would be due to the continuation of the process of decomposition of the substance after the stimulus has ceased. The negative after-image would be due to a decrease in one of the substances (the one corresponding to the color of the stimulus) while the other two are found in normal quantities; white light would, therefore, provoke a greater response from the latter and a smaller response from the former. However, not all the facts observed are explained by this theory. It gives no explanation of the sensations of gray and white, which arise in the peripheral retina where there is no sensitiveness for color, or for the sensation of yellow given by stimulation of an area outside the part of the retina sensitive to red and green (according to the theory, yellow is the result of combined stimulation by red and green). Finally, it has no satisfactory explanation for total color blindness with persistence of the vision of form. Observations made on the electrical phenomena of the retina and optic nerves by Granit have brought forward further evidence in favor of this theory (see Chap. 78).

Hering's theory. According to Hering there are three photosensitive substances, which on decom-

position or resynthesis provoke different sensations by stimulation of different nerve fibers. These sensations are those of red-green, yellow-blue, and white-black. White light breaks down the substance sensitive to white, which on being resynthesized provokes impulses which produce the sensation of black. The existence of this substance sensitive to white would explain vision in cases of total color blindness. According to this theory complementary colors (*e.g.*, red and green or blue and yellow) have opposite and antagonistic effects.

Several other theories of color vision have been proposed. Motokawa and his associates,[1] observing the effects of electric stimulation of the eye, conclude there are four different types of receptors. Four principal photosensitive substances have been identified (see Chap. 78).

Disturbances in Color Vision

Dalton, the English physicist, discovered color blindness in himself and in his brother, and he gave a classic description of the anomaly in 1794. Blindness to one or more colors is frequently observed. According to certain statistics 8 per cent of otherwise normal males and 0.5 per cent of females are color-blind. It is a hereditary sex-linked character; the defective factor responsible for its appearance is found in the X chromosome, and hereditary transmission takes place in the same way as for hemophilia. A color-blind person often does not realize his condition or may discover it only late in life. There are many simple tests that quickly reveal its existence. Color blindness is of importance in persons such as engine drivers, pilots, etc., who are guided by colored traffic lights, as they may unknowingly commit errors with disastrous consequences. Normal color vision is also necessary in other professions and trades, *e.g.*, painting, dyeing, printing, color photography and television, horticulture, etc.

Classification of color blindness. Usually a color-blind person is said to be blind to a given color, *e.g.*, red, green, etc. This is not strictly true, because there is not always total blindness to a color, and the defect, although it is predominant for one color, always extends to others. The classification now used was proposed by von Kries and is based on the Young-Helmholtz theory of the existence of three specific color receptors. According to this classification, normal vision is trichromatic; deficiency in color vision may affect one, two, or all three types of color receptors.

Trichromats	*Dichromats*	*Monochromats*
Normal	Protanopia	
Protanomalous	Deuteranopia	
Deuteranomalous	Tritanopia	

Anomalous trichromatic vision. Individuals with this type of vision see the three primary colors and synthesize white and the other colors of the spectrum by fusion of the primary colors. They are therefore trichromats. They perceive red and green differently from normal subjects, although they are not blind to these colors. In mixing red and green to form yellow, they use proportions different from those chosen by normal subjects. Protanomalous individuals are deficient in vision of red (they require more red), and deuteranomalous ones are deficient in vision of green.

Dichromatic vision. Individuals with this type of vision do not see one of the primary colors, supposedly because the corresponding receptor is missing. They see the two other primary colors, although with certain deficiencies. They learn, however, to recognize the color of usual objects from their brightness and from previous experience. The diagnosis can be made because these subjects attempt to synthesize white and all the colors of the spectrum by combining only two colors.

In protanopia, the most common anomaly, there is blindness to red; green is deficiently perceived, and blue is seen normally. In deuteranopia there is blindness to green; red is deficiently perceived, and blue is seen normally. In tritanopia there is blindness to blue, but red and green are perceived. Tritanopia is much less frequent as a congenital anomaly; about 1 in 65,000 persons suffer from it.[1] Jaundice, intoxication by santonin, and dislocation of the retina may provoke it. According to König, there is normally a small tritanope point in the center of the fovea, which can be easily demonstrated.

Monochromatic vision. Achromatopsia or total color blindness. Individuals with monochromatic vision do not perceive color; they see only white, black, and different shades of gray. They distinguish colors exclusively by differences in brightness.

[1] Ebe, M., K. Isobe, and K. Motokawa, *Science*, **113**, 353, 1951.

[1] Wright, W. D., *Nature, London*, **170**, 904, 1952.

Night vision is normal. In these individuals apparently there is only rod vision; the cones do not function. Achromatopsia is an extremely rare congenital anomaly.

The spectrum in color blindness. Subjects with anomalous color vision see the scotopic spectrum (*i.e.*, in the dark) normally. The daylight or photopic spectrum is seen normally by deuteranomalous subjects and in deuteranopia. For protanomalous subjects and in protanopia, the spectrum is shortened at the red end and the brightest part is in the blue wavelengths (540 mμ), instead of in the yellow (554 mμ).

Tests of color vision. There are many tests of color vision. In the Holmgren test the subject is given a skein of colored wool and told to choose from skeins of assorted colors those of similar hues. Another test consists in showing lights of different colors and intensity which the subject must recognize. Ishihara's test (a modification of Stilling's test) consists of a series of plates in which a digit formed of spots of one color is "hidden" in a field of other colored spots. Subjects suffering from different forms of color blindness do not see the digit. By an adequate choice of colors and shades the test can be made diagnostic for the different forms and intensity of color blindness.

The most accurate method consists in asking the subject to match spectral colors; *e.g.*, to match yellow by mixing red and green. The proportions of each color used are compared with those used by normal subjects. The range of the visible spectrum and the point of maximum brightness in the spectrum are also determined, and the subject is asked to discriminate between different spectral hues.

BIBLIOGRAPHY

See page 976.

CHAPTER 77

The Physiology of the Eyeball and Its Appendages

THE EYEBALL IS a complex structure formed by closely integrated parts, each one fulfilling a special function. The eye is protected by the secretion of the lacrimal glands and the eyelids, and it is moved by the extrinsic muscles of the eye. The physiology of the cornea, iris, crystalline lens, and ocular fluids will also be considered in this chapter; the retina and the optic pathway will be dealt with in the following chapter.

LACRIMAL SECRETION

The secretion of tears together with the eyelids forms a protective mechanism for the eyeball.

Anatomy. There is a principal orbital gland situated in the upper external part of the orbital cavity, and there are several small glands distributed along the fundus of the conjunctiva. Histological examination shows that they are formed by acini of secretory cells, the ducts of which converge to one or more excretory ducts opening into the fundus of the conjunctiva. The cells show different phases of secretion, with formation and excretion of granules.

Normal lacrimal secretion. It is not possible to measure accurately the total amount secreted during the day. Schirmer (1904) calculated that 13 drops are excreted every 16 hr., of which 7 drops are evaporated and the rest flow into the nasal cavities.

Irritation of the conjunctiva or the nasal mucosa increases lacrimal secretion. The afferent fibers of this reflex form part of the trigeminal nerve. Vomiting, sneezing, and coughing are accompanied by increased lacrimal secretion. In man and the anthropoids, lacrimal secretion increases in certain psychic states. Weeping forms part of the expression of sorrow and extreme hilarity. It can also be provoked by a conditioned reflex, *e.g.*, when actors simulate painful emotions.

Mechanism of lacrimal secretion. Tears are secreted by a mechanism similar to that of salivary secretion. Sympathetic and parasympathetic nerves play a fundamental part in this process. The glands are innervated by parasympathetic secretory and vasodilator fibers which arise in a nucleus of small motor cells situated in the pons near the upper salivary nucleus. These fibers form part of the seventh cranial (facial) nerve, from which they diverge at the level of the geniculate ganglion. They end in the sphenopalatine ganglion. Postganglionic fibers from this ganglion travel to the lacrimal gland in the maxillary division of the trigeminal nerve. Parasympathomimetic drugs such as pilocarpine, muscarine, and acetylcholine stimulate lacrimal secretion, which is inhibited by parasympatholytic drugs, such as atropine.

Sympathetic fibers arise in the intermediolateral cell column of the upper thoracic segments of the spinal cord. The preganglionic fibers travel in the cervical sympathetic and end in the superior cervical ganglion. Postganglionic fibers arising in this ganglion ascend in the carotid plexus. They provoke vasoconstriction and have a secretory effect which is less marked than that of parasympathetic fibers.

TEARS

The chemical composition of tears is given in Table 100. Lysozyme has been found in tears; it

has a bacteriostatic effect which prevents the growth of bacteria. Lysozyme concentration diminishes in cases of eye infection and again increases when the condition of the patient improves.

Table 100. Chemical Composition of Tears

Constituent	*Per Cent*
Water	98.2
Solids	1.8
Total proteins	0.67
Chloride	0.66
Glucose	0.65
Sodium	0.44
Potassium	0.12

Source: Ridley, 1930.

Tears flow to the inner angle of the eye, where they are collected in the lacrimal sac and pass through the lacrimal canals into the nasal cavity. Contraction of the orbicularis plays an important part in the circulation of tears. Paralysis of this muscle causes the tears to flow down the cheek. Normally the orbicularis, on contracting, compresses and empties the lacrimal sac, but the sac can be extirpated without disturbing the circulation of tears, which continue to flow by capillarity into the lacrimal canals.

THE EYELIDS

The movements of the eyelids, by closing and opening the eyes, regulate the entrance of light into the eye. These movements protect the eyes from excessive illumination and the action of foreign bodies. They also spread the tears secreted by the lacrimal gland over the exposed part of the eyeball, thus preventing desiccation of the cornea and conjunctiva.

Neuromuscular apparatus of the eyelids. The eyelids have two muscles, the orbicularis and the levator palpebrae. The former closes the eyelids; the latter raises the upper lid. These muscles are reciprocal antagonists; when one contracts, the other relaxes. The postural tone of these muscles varies considerably. Relaxation of the levator palpebrae is a sign of sleepiness. During sleep the tone of the orbicularis is predominant and the eyes are closed. In the waking state the levator is tonically contracted and it keeps the eye open.

The orbicularis is innervated by the facial nerve and by sympathetic fibers, and the levator by the oculomotor nerve and parasympathetic fibers. The nerve centers of the sympathetic and parasympathetic fibers are situated in the hypothalamus. The facial nuclei and the oculomotor nuclei give rise to the motor fibers for the orbicularis and the levator palpebrae. Cortical centers for the movements of the eyelids are found in the lower part of the precentral gyrus.

MOVEMENTS OF THE EYELIDS

Movements of the eyelids are associated with movements of the head and eyes: (*a*) certain movements are accompanied by movements of the eyelid, *e.g.*, when the head and eyes are lifted upward, the upper lids are raised; (*b*) movements of the eyelids are associated with other movements, *e.g.*, the pupil contracts and the eyeball is rotated upward and outward on closing the eyelids; (*c*) the movements of one eyelid are usually associated with those of the other eyelid, and they also take an important part in facial expression.

There are three principal movements of the eyelids: voluntary blinking, reflex blinking, and spontaneous blinking.

Voluntary blinking. The eyes cannot be kept open without blinking for more than a few minutes, but they can be kept voluntarily closed for an indefinite time. One eye alone can be closed while keeping the other open. Voluntary blinking is a very rapid movement. It takes only 0.2 sec., about one-half the time taken by spontaneous blinking.

Reflex blinking. A large number of stimuli, mainly acting on receptors innervated by the second (optic), third (trigeminal), and eighth (acoustic) cranial nerves, provoke defensive reflex closure of the eyelids. For example a bright light that suddenly illuminates the eye, a loud noise, or physical contact with the cornea, conjunctiva, eyelashes, or the nasal mucous membrane cause reflex blinking.

Spontaneous blinking. This is the most frequent movement of the eyelids. It consists in rhythmic involuntary closing and opening, at a rate of 10 to 20 movements per minute. The tears are thus spread over the eye, and desiccation of the cornea and conjunctiva is prevented. The movement consists of very rapid closing, followed by the opening of the eye. The whole movement lasts 0.3 to 0.4 sec., and it is repeated after an interval of 2 to 10 sec. Slow-camera cinematography shows that the upper lid falls like a curtain; the lower lid remains almost motionless. The eye closes progressively from the outer to the inner angle, sweeping tears and

foreign bodies toward the lacrimal sac. Lawson[1] has pointed out that owing to blinking there is a blackout during 10 per cent of the time the eyes are active, which increases if the rate of blinking is greater. This causes a certain degree of inaccuracy in the observation of very rapid phenomena and also in controlling movements where speed is essential (*e.g.*, when playing tennis or boxing) or in handling very fast machines; *e.g.*, a jet-propelled plane travels several hundred yards in the time taken by a single blink.

The cause of spontaneous blinking is still unknown. It is not due to drying of the cornea, because it persists after the eye has been anesthetized or in an atmosphere saturated with water vapor in which the cornea is not dried. Neither is blinking due to light reflexes, because it persists in the dark and in blind persons. Perhaps it is a reflex arising in proprioceptive receptors in the eye muscle, or it may be due to rhythmic automatic discharge of the nerve centers.

THE MOVEMENTS OF THE EYE

Each eye performs several fast and precise movements which are closely associated with the movements of the other eye. Any eye movement involves at least two muscles of each eye. These movements are controlled by accurately integrated nerve impulses.

Anatomy. Six extrinsic muscles move each eyeball.

The four recti muscles are inserted on a fibrous ring (tendon of Zinn) attached to the margin of the optic foramen. Their fibers are directed forward, opening out, so as to form the four sides of a pyramid. The *external and internal recti* end near the margin of the cornea on a flat tendon; the lines of insertion of these tendons are parallel to the corneal margin. The *superior and inferior recti* end on tendons that are inserted lateral to the middle axis of the eye, behind the corneal margin; the lines of insertion are not parallel to the corneal limbus.

The superior oblique arises in the depth of the orbit and passes to its upper and inner side. There the muscle forms a tendon which passes through a fibrous ring, acting like a pulley. The direction of the fibers of the muscle changes, and they pass backward and downward under the superior rectus and end on the superior posterior and external aspect of the eyeball.

[1] Lawson, R. W., *Nature,* **161,** 154, 1948.

The inferior oblique arises in the inferior and internal angle of the orbit and passes outward, under the eyeball, ending on its inferior posterior and external aspect.

The oculomotor (third cranial) nerve innervates all the extrinsic muscles of the eyes except the external rectus, which is innervated by the sixth cranial nerve (n. abducens), and the superior oblique, which is innervated by the fourth cranial nerve (n. trochlearis). The nuclei of the third and fourth nerves are found in the midbrain. The oculomotor nucleus consists of a series of nuclei, one for each muscle. The nucleus of the sixth nerve is found in the pons.

Movements of the Eyeball

Since the middle of the nineteenth century many methods have been applied to the study of the movements of the eyeball, *e.g.*, direct observation of the eyeball, photographic or cinematographic registration of its movements, and the mechanical and electrical registration of the contraction of the different muscles.

The eyes at rest. Fixation of the eyeball. Absolute rest is obtained only by section of all the nerves of the eye muscles; the eyes then look forward in a slightly divergent outward direction. The eyeball is physiologically in the resting position when it is not submitted to any stimulus that provokes accommodation or fusion of images, but it is still under the control of muscles with their innervation intact. In this position the eyes look forward, and their axes are parallel. When the eyes are fixed on an object, both are displaced from the physiologic resting position so that the images are formed on corresponding points of the retinas of both eyes. They are kept in that position by a sustained contraction of the eye muscles made up of rapid, asynchronous twitches of their fibers.

Rotation and torsion. All the movements of the eye are rotations around several axes. The eyeball behaves as if it were a condyle in a condyloid articulation, the socket formed by the orbital fat. The theoretical center of rotation of the eyeball is placed 14 mm. behind the center of the cornea. A frontal plane passing through the center of rotation (Listing's plane) is used as a reference for the primary axes of rotation, *i.e.*, the vertical and horizontal axes. Rotation around the vertical axis moves the cornea to the right and to the left. Rotation

around the transverse axis moves the cornea up and down. These two types of rotation are known as cardinal movements and are the only pure voluntary movements. Rotation around the horizontal (sagittal) axis, *i.e.*, the line passing through the center of rotation and the object

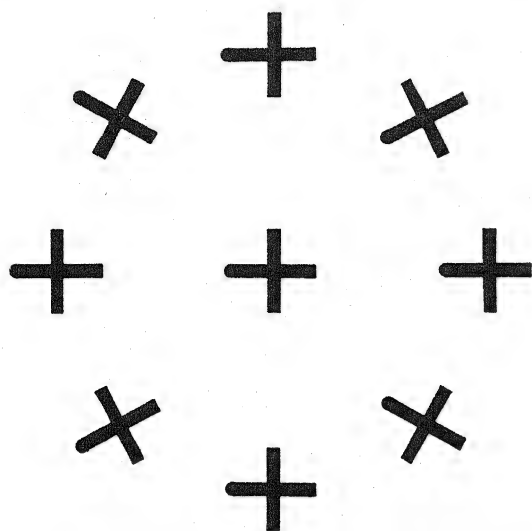



FIG. 423. Demonstration of torsion movements of the eyeball by means of the after-image. (*Ruete.*)

(fixation point) produces wheellike movements of the cornea. Rotation around the oblique axes, which pass through the center of rotation at oblique angles to the horizontal axis, produces involuntary torsion movements that cause the cornea to rotate like a wheel. These movements can be easily followed by tattooing a point near the corneal margin. Torsion movements can be appreciated subjectively in the following way: A bright red cross is placed on a gray background, and the eyes are fixed on it for 20 sec. or more. The eyes are then quickly turned to the right or left, or up or down (cardinal movements); the after-image is always that of a vertical cross. If oblique movements are made, by looking at an angle, the cross of the after-image appears to have rotated more or less to one side, owing to the torsion of the eyeball (Fig. 423). Donders, Helmholtz, and Listing have made detailed studies of torsion movements and expressed their results in the laws that bear their names.[1]

Movements of the Extrinsic Muscles of the Eye

The eyeballs are moved by striated muscles which have a very short contraction time and a permanent postural tone that fixes the eyeball. Reciprocal innervation is prominent in the functioning of these muscles. Relaxation of the antagonist on contraction of the protagonist can be easily demonstrated as follows: All the extrinsic muscles of one eye are cut except the external rectus. The animal is shown an object, which is then displaced toward the opposite side. The eyeball follows the object until the pupil is situated in the mid-line, *i.e.*, in the resting position. This movement is due exclusively to relaxation of the external rectus, all the other muscles (including the internal rectus) having been cut.

The central nuclei of these muscles situated in the pons and brain stem are joined together and to the nuclei on the opposite side by many association fibers, which assure a closely coordinated action. The oculomotor nuclei are also connected to the vestibular nuclei and cortical centers in the second frontal convolution and the occipital lobe. These cortical centers govern the associated movements of the head and eyes; if they are stimulated, conjugate movements of the eye are provoked.

Individual action of each muscle. The isolated contraction of an extrinsic muscle causes the eyeball to rotate around the center of rotation and the cornea is displaced. If the lateral recti contract, the cornea moves horizontally to the right or to the left according to which muscle has contracted. If any one of the four other muscles contracts, the cornea undergoes a triple displacement, because these muscles are inserted obliquely with respect to the axis of the eyeball. Thus the superior rectus causes not only an upward movement but also a slight inward displacement and torsion of the eyeball. When the oblique muscles contract the main displacement is one of torsion. In Fig. 424 the movements of the eyes are diagrammatically represented.

Associated movements. The movements of the eyes are normally the result of the simultaneous contraction of two or more muscles. Associated contraction results in the reciprocal annulment of opposite effects. It can be seen

[1] Donders' law (1847) states that for every position of the visual axis with respect to the head, there is a definite and constant angle of torsion independent of the observer's will and of the manner in which the eyes are focused on the fixation point.

in Fig. 424 that the eye can be directed outward (*a*) by contraction of the external rectus; (*b*) by the associated contraction of the superior and inferior oblique muscles; (*c*) by associated contraction of the three muscles.

Bilateral Movements

The movements of both eyes are coordinated so that they are "symmetrical and equal," as Hering stated in 1879. They are moved reflexly or voluntarily.

uniform; the eyes are rotated and fixed alternately. In reading, for example, a group of letters is fixed and after a short pause (0.2 sec.) the eyes are rapidly moved (0.025 sec.) to the next group.

Coordination of voluntary movements is a complicated process. Electrical stimulation of the posterior part of the second frontal gyrus provokes conjugate movements of the eyes toward the opposite side. Destruction of these cortical centers does not cause paralysis of the eye muscles, but conjugate movements are lost.

Fig. 424. Movements of the cornea produced by the contraction of each one of the eye muscles, according to Terrien. The line indicates the inclination of the vertical meridian (torsion).

Voluntary eye movements. When the eyes are directed upward or downward, to the right or to the left, the visual axes are always parallel to each other and the movements are known as conjugate. The area of fixation is limited by the maximum range of these movements made in all directions, without moving the head. It is determined by keeping the head fixed and following an object which is displaced in the visual field in all directions and as far as it can be clearly distinguished. The summated area of both eyes is given by a rotation of approximately 40° in an upward direction, 60° downward, and 50° outward.

When the visual axes of both eyes are not parallel but converge to a point, the movements are known as "convergent." They are called "divergent" or "disjunctive" movements when the visual axes pass from a convergent to a parallel direction. Both eyes take part in convergent movements even when the point of fixation is situated on the visual axis of one eye; this eye undergoes slight oscillatory and torsion movements. If the fixation point is approximated to the observer, convergence increases. It is at a maximum at 8 cm. from the eye (near point); at lesser distances a double image is formed.

When the movements of the eye are registered photographically or cinematographically, it is seen that the movements are not smoothly

Reflex eye movements. Reflex movements of the eye are involuntary but are accompanied by a certain degree of consciousness.

The reflex movements for fusion of the images prevent the formation of blurred images by directing the visual axes so that the images will be formed on corresponding points of the retina. For example, if a weak prism is placed before one eye at first there is diplopia, but soon a correcting reflex develops which assures the formation of a single image.

Fixation reflexes, which rapidly focus on the fovea an image formed on the peripheral part of the retina, are also accompanied by a certain degree of consciousness.

Postural reflexes of the eye are outside the sphere of consciousness. They contribute to the maintenance of bodily equilibrium. Stimulation of certain proprioceptors provokes reflex eye movements such as nystagmus (see Chap. 81).

Disturbances in eye movements. In certain cases movements of both eyes are not harmoniously combined; convergence to a point of both visual axes may be impossible (strabismus) or obtained only with effort (heterophoria). In both cases muscular synergy does not exist. Visual images are not formed on corresponding parts of both retinas, and there is *diplopia, i.e.,* two images of the object are perceived. In strabismus one of the images is suppressed by a psychologic process.

THE CORNEA

The cornea is a membrane covering the anterior part of the eyeball; it is a continuation of the sclera. It has no blood vessels. It is transparent and acts as a lens of 41 to 45 D., because it is placed between the air and the aqueous humor. It has a special type of sensibility.

Anatomy and histology. The cornea is like a watch glass fixed in the sclera. Its diameter is 11.6 mm., and its thickness 0.8 mm. in the center and 1 mm. in the periphery. The anterior aspect has a curvature with a radius of 7.84 mm. and the posterior aspect a curvature with a radius of 6.8 mm. In the adult it weighs 180 mg.

Microscopical examination shows that the central or optical part differs from the peripheral part or limbus. The optical part is formed by five layers: (*a*) stratified epithelium, 50 to 100 μ thick, continuous with the epithelium of the conjunctiva; (*b*) Bowman's anterior lamina, which is structureless and 10 μ in thickness; (*c*) the substantia propria of the cornea, consisting of fibers arranged in parallel rows forming approximately 60 laminae separated by cell spaces or lacunae (it forms 60 per cent of the whole corneal thickness); (*d*) the posterior lamina of Descemet with a thickness of 5 to 7 μ; (*e*) a single layer of flat endothelial cells, which is in contact with the aqueous humor. The peripheral part of the cornea is transitional to the sclera. It has large arterial and venous loops, lymphatic vessels, and nerve endings.

The epithelium of the cornea is covered in the living eye by a thin fluid film, which can be seen with the ocular microscope, especially when fluorescein has been instilled in the eye. This precorneal film has its origin in lacrimal and conjunctival secretions; it is a colloid dispersion which is being continuously renewed. It protects the corneal epithelium and dissolves drugs which after instillation penetrate into the eyeball.

Chemical Composition

The chemical composition of the cornea is given in Table 101. The water decreases with age, so that the other components of corneal tissue become more concentrated. The ash contains Cl, K, Na, Mg, Ca, sulfate, and phosphate. Fats are neutral fats, phospholipids, and cholesterol. Protein forms 98 per cent of the solids. There is a mucoprotein which has mucoitin–sulfuric acid and is similar to the mucoproteins in the vitreous humor and in the umbilical cord. There is also collagen, which gives gelatin on boiling, and a small amount of elastin. Radioactive iodine injected into the organism is found in greater proportion in the cornea than in other eye tissues.

A relatively large amount of acetylcholine has been found in the cornea. It diminishes after the trigeminal nerve has been cut, in proportion to the degree of ulceration of the cornea following the operation.

Table 101. Chemical Composition of Ox Cornea

Constituent	*Per Cent*
Water	81.13
Solids	18.87
Protein	
Total	18.53
Mucoid	19
Collagen	81
Elastin	trace
Ash	0.17
Lipids	0.04

Source: Krause, 1934.

In certain pathologic conditions, urates, hemosiderin, and other foreign substances are deposited in the cornea.

Physical Properties

Hydration. When the cornea is placed in water, it swells and increases up to five times its normal weight (imbibition or turgescence). The sclera does not have this property, although its histologic structure is similar to that of the cornea. Moreover, the cornea does not take up water from the aqueous humor *in vivo*, but swells when placed in aqueous humor *in vitro*.

Permeability. The permeability of the cornea *in vivo* is shown by the rapidity with which substances in solution placed on the cornea are absorbed into the eye and rapidly produce their effects. Cogan and Kinsey[1] have shown that *in vitro* water can flow through the cornea in both directions without producing edema. The posterior endothelial and anterior epithelial surfaces behave as semipermeable membranes. Rupture of either of these membranes is followed by edema. The normal epithelium is impermeable to NaCl in both directions.

Transparency. The transparency of the normal cornea *in vivo* is apparently perfect. However, if it is examined with an ocular microscope it appears to be gray because of the reflection

[1] Cogan, D., and E. V. Kinsey, *Arch. Ophth.*, **27**, 466 and 696, 1942; **28**, 272, 449, and 661, 1943; **31**, 408, 1944.

of light by its parenchyma. In the 5 months' fetus it is opaque, and it becomes opaque a few hours after death.

The transparency of the cornea is dependent on the water it contains; if it is dehydrated it becomes opaque and regains its transparency on rehydration. The sclera has a microscopic structure similar to that of the cornea, and on being hydrated it becomes transparent.

Cornea
Anterior chamber
Lens

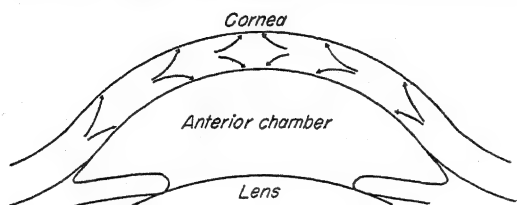



FIG. 425. Diagram of circulation of fluids in the cornea. (*Kinsey and Cogan.*)

According to Kinsey and Cogan, the normal transparency of the cornea depends on an equilibrium between the processes of hydration and dehydration. Hydration is due to the continuous flow of fluid from the lymph vessels in the corneal limbus. Dehydration is due to the fact that the anterior and posterior aspects of the cornea are in contact with hypertonic fluids (tears and aqueous humor), which draw water from the cornea through the semipermeable membranes that line it. The surface of hydration in the corneal limbus measures 0.4 sq. cm. and that of dehydration 2 sq. cm. If the normal equilibrium between the processes is disturbed, the cornea becomes opaque (Fig. 425).

Reflection of light. Light rays are partially reflected when they pass from one medium to another with different optic density. This occurs on the anterior and posterior aspects of the cornea and is the origin of the well-known phenomenon called "Purkinje's images," which will be considered in the discussion of the lens. The amount of light reflected by the different surfaces, according to Raeder (1922), is as follows: anterior surface of the cornea, 2.5 per cent; posterior surface of the cornea, 0.024 per cent; anterior surface of the lens, 0.036 per cent; posterior surface of the lens, 0.030 per cent. Images reflected from the anterior surface of the cornea are therefore approximately 100 times as bright as those from the posterior surface, and 70 times as bright as those from the surface of the lens.

By means of the ophthalmometer (invented by Helmholtz and later considerably improved) it is possible to measure the curvature and determine the regularity of the anterior and posterior surfaces of the cornea, and also to measure its thickness.

PHYSIOLOGY

Circulation. In the normal adult cornea the blood vessels do not penetrate beyond the limbus. Section or destruction of half the limbus does not provoke opacity of the cornea, but if the limbus is totally destroyed the cornea loses its transparency. Fluid passes out of the capillaries in the limbus and seeps through the lacunar spaces between the laminae of the cornea, not through special ducts. This centripetal flow can be demonstrated in the following way: The cornea is pricked with a rusty needle in several places, and potassium ferricyanide is injected intravenously. The peripheral points are stained blue before those placed nearer the center. This demonstration has been confirmed by injecting radioactive ions and following their appearance and movements in the cornea.

Metabolism. Friedenwald[1] has recently made an important contribution to the knowledge of the metabolism of the cornea. The isolated bovine cornea takes up oxygen at the rate of 70 cc. per hr.; 90 per cent of the oxygen consumption is due to the activity of the epithelium. The rate of oxygen consumption varies in different species.[2] CO_2 is given off, and the RQ is 1.

Glucose is completely oxidized by the epithelium, but the substantia propria converts glucose into lactate or pyruvate, which is afterward oxidized by the epithelium.

Sensibility. The sensibility of the cornea is due to the many nerve fibers that end in it. They can be observed *in vivo* by means of the ocular microscope. If the nerves in the limbus are cut, the cornea loses its sensibility for a time, but soon recovers it because the nerves regenerate fairly rapidly. When a corneal graft is made, the fragment grafted becomes sensitive in approximately a year.

These fibers make the cornea sensitive to pain, temperature, and touch, and give rise to reflexes such as the corneal reflex, which provokes closure of the eyelids. Sensibility to pain is more

[1] FRIEDENWALD, J., *Bull. Johns Hopkins Hosp.*, **82**, 1948 (the whole volume).

[2] EPSTEIN, M., *Vest. Oftal.*, **29**, 22, 1950; DE ROETTH, A., *Arch. Ophth.*, **44**, 666, 1950; **45**, 139, 1951.

developed in the central part of the cornea than in the peripheral parts. Each one of the fibers penetrating from the periphery innervates a quadrant of the cornea, which in part overlaps the area covered by neighboring fibers.

The peripheral part of the cornea is sensitive to cold and touch. There is still some discussion as to whether the center also gives rise to tactile as well as painful sensations. Franceschetti[1] has observed that after anesthetizing the cornea there are two thresholds of recovery, which he interprets as due to the reappearance of tactile and painful sensations respectively. Local application of cocaine or cocaine derivatives rapidly produces anesthesia of the cornea but causes some damage to the epithelium.

THE IRIS AND PUPIL

The Iris

The iris is a pigmented disklike membrane placed in the anterior part of the eye. It is perforated in the center, the orifice being called the pupil. The size of the pupil varies according to the state of contraction or relaxation of the muscle fibers of the iris, thus regulating the amount of light that enters the eye. The aqueous humor is also formed at the level of the iris, and the reticuloendothelial cells in the iris take up cell debris and foreign bodies in the aqueous humor. The pigment in the iris prevents the penetration of light rays into the eye except through the pupil.

The Pupil

In man the form of the pupil is circular and it is placed almost in the center of the iris, but its form varies in different species. In the shark and other fishes it is oval; in saurians it is a vertical slit with dentate borders, which interlock when the pupil contracts. Among the mammals it is oval in the horse and sheep, a vertical slit in the cat, and a horizontal slit in the goat.

The size of the pupil varies according to the physiologic circumstances. Contraction of the pupil is known as *myosis* and dilatation as *mydriasis*. In women the pupils are usually larger than in men; in children and old people they are often small. In young men of twenty the average diameter is 4 mm., which is reduced to 2.5 mm. between the ages of forty and fifty.

[1] Franceschetti, A., *Ber. über die XLIX Zusam. der Deuts. Ophth. Ges.*, Leipzig, p. 242, 1932.

The pupil is contracted to a diameter of 1 mm. in a bright light, and dilates to 8 mm. in poor light. In emotional states such as fear, the pupil dilates. It contracts during natural sleep and narcosis produced by anesthetics; simulation of sleep can thus be detected. Immediately after death there is marked mydriasis, but after a time the diameter of the pupil diminishes.

Movements of the Iris

Changes in the diameter of the pupil are due to the contraction and relaxation of two muscles in the iris, the sphincter and the dilator. These muscles arise from the pigmented epithelial layer of the retina. They are smooth muscles and are not under voluntary control.

The sphincter is the stronger of the two. It forms a ring 0.8 mm. thick around the pupil. On contraction its fibers can shorten their length as much as 87 per cent. Such intense contraction is not observed in any other muscle in the organism.

The dilator is formed by radial fibers, which go from the ciliary muscle to the margin of the pupil. These are myoepithelial fibers which have not been completely transformed into muscle fibers. They had passed unnoticed until Langley and Anderson observed local retraction of the margin of the pupil when the iris was stimulated electrically at different points. This observation led to the search for the fibers responsible for this movement and the discovery of the dilator muscle.

Innervation

The pupil is innervated by fibers of the two divisions of the visceral nervous system. The sympathetic innervates the dilator, and the parasympathetic the sphincter. Centers of pupillary activity have been localized in the hypothalamus and the spinal cord, and stimulation of the frontal eye fields of the cortex (area $8\alpha\beta\delta$) is occasionally accompanied by mydriasis.

Sympathetic. Karplus and Kreidl discovered a center in the hypothalamus that controls the movements of the iris, and Budge described a pupillodilator (the so-called "ciliospinal center") in the spinal cord, almost a century ago. This center is situated in the intermediolateral cell column of the spinal cord from the last one or two cervical segments to the first three or four thoracic segments. The preganglionic fibers emitted from this center leave the spinal cord

in the white rami of the corresponding ventral roots and travel up the cervical sympathetic to the superior cervical ganglion, where they end. The postganglionic fibers form part of the trigeminal and long ciliary nerves and end on the fibers of the iris (Fig. 426).

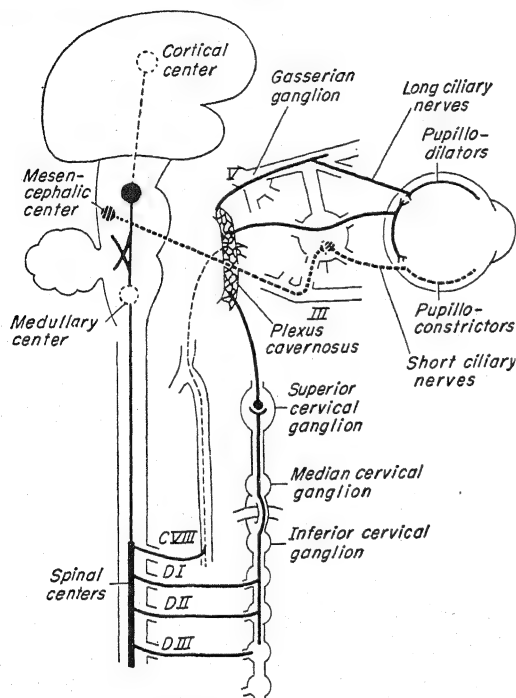



Fig. 426. Diagram of innervation of the pupil.

In the course of his classic observations on the cervical sympathetic, Claude Bernard described the effects produced on the eye by section of this nerve in the rabbit, *i.e.*, myosis, enophthalmia, narrowing of the palpebral fissure, and vasodilatation. Stimulation of the cervical sympathetic produces the opposite effects, among which is pupillary dilatation (mydriasis), due to the contraction of the dilator.

Parasympathetic. Contraction of the sphincter causing myosis is governed by parasympathetic fibers. Section of these fibers causes dilatation of the pupil; they are, therefore, antagonistic to the sympathetic fibers.

There is a hypothalamic vasoconstrictor center, and situated close to the oculomotor nucleus in the brain stem is the well-known Edinger-Westphal nucleus. Fibers from the latter emerge together with the fibers of the oculomotor nerve and end in the ciliary ganglion. The postganglionic fibers form part of the short ciliary nerves and end on the fibers of the sphincter (Fig. 426).

Pharmacology

Several drugs modify the diameter of the pupil (Table 102). They act by stimulating the effectors of the sympathetic or parasympathetic or paralyzing one or the other division of the autonomic nervous system, thus disturbing the equilibrium that normally exists between the two divisions (see Chap. 84).

Table 102. Effect of Drugs on the Pupil

Drug	*Sphincter*	*Dilator*
Mydriatic drugs		
Sympathomimetic		
Adrenaline	....	Contracts
Ephedrine	....	Contracts
Benzedrine	....	Contracts
Parasympatholytic		
Atropine	Relaxes	
Scopolamine	Relaxes	
Homatropine	Relaxes	
Myotic drugs		
Parasympathomimetic		
Choline and derivates	Contracts	
Pilocarpine	Contracts	
Muscarine	Contracts	
Eserine	Contracts	
Sympatholytic		
Ergotamine	....	Relaxes
Ergotoxin	....	Relaxes

Pupillary Reflexes

The diameter of the pupil is modified reflexly by stimuli that act on several receptors; there are also changes in the size of the pupil associated with other eye movements. Strong stimuli such as a foreign body causing damage to the cornea or conjunctiva, a very bright light, a loud noise, etc., cause mydriasis. The most important of the pupillary reflexes is the light reflex.

The light reflex. Light falling on the retina of one eye causes contraction (myosis) of the pupils of both eyes. The stimulus is white or monochromatic light, which must reach a certain threshold. The minimal strength of the stimulus increases as the point stimulated is further from the fovea. It must also last a certain time and be produced with relative suddenness.

The receptor is the retina, and the light reflex is abolished by destruction of the retina. Dark and light adaptation modify the sensibility of the retina to light with respect to this reflex. The threshold is lowered proportionally to the area of the retina stimulated. After repeated application of the stimulus, the response diminishes owing to fatigue.

Ranson and his associates have made important contributions to the knowledge of the path followed by the impulses that provoke this reflex (Fig. 427). The centripetal impulses travel along the optic nerve, chiasma, and optic tract to the pretectal region, where they have their first synapse. Fibers from this region cross over to the other side in the posterior commissure and end in the Edinger-Westphal nucleus. Efferent fibers leave this nucleus and reach the iris through the oculomotor and short ciliary nerves. Ranson's work has shown that the path of the light reflex does not pass through the lateral geniculate body or through the colliculi. Visual sensations and the light reflex do not have a common path, and it is still doubtful whether they have a common afferent fiber that branches out, or whether there are separate sensory fibers for each kind of impulse. The latter supposition is based on the fact that there are cases of blindness due to injury of the optic nerve in which the light reflex is not abolished.

FIG. 427. Nerve path of light reflex.

The response takes place in the pupil of the illuminated eye (direct light reflex) and of the opposite eye (consensual light reflex). There is at first a rapid constriction of the pupil, which then gradually dilates to the habitual size, as the retina becomes adapted to the higher level of illumination. The latent period of the initial constriction is 200 to 500 msec., considerably greater than the latent period of the knee jerk (60 msec.); therefore the reflex path is obviously made up of several neurons.

Associated movements. When the eyes are fixed on an object nearer than infinity (*i.e.*, within 20 ft.) the following movements take place: (*a*) the visual axes of both eyes converge; (*b*) the lens undergoes accommodation; (*c*) the pupil is constricted. Myosis is bilateral and dependent on accommodation and convergence, mainly on convergence; but either accommodation or convergence can be suppressed by appropriate lenses or prisms, and myosis is then related solely to one or the other process.

Myosis is also associated with reflex or voluntary closure of the eyelids.

THE CRYSTALLINE LENS

The lens is biconvex and is placed in the path of the light rays that enter the eye. It is elastic, and its shape can be changed so as to focus the image on the retina, whether the rays are emitted from an object placed at infinity or from one near the eye. The lens is thus the most important dioptric medium of the eye; it assures the formation of a clear image on the retina at whatever distance from the eye the object is placed.

Anatomy. The lens is a transparent and elastic structure which has no blood vessels or nerves. It is suspended between the posterior chamber and the vitreous humor by the zonula, a suspensory ligament formed by fibers attached on the one hand to the ciliary body and on the other to the margin and the neighboring parts of the anterior and posterior surfaces of the lens.

The lens is biconvex; its posterior surface has a

greater curvature than the anterior surface. It has a slight yellow tint in young subjects, which becomes more marked in old age. Its weight increases gradually throughout life; at thirty years of age it weighs approximately 180 mg.

Histological examination of the lens shows that it is formed by concentric layers of fibrous cells, covered by a thin, elastic, permeable membrane. The central nucleus is formed by the older fibers. With age it loses water and hardens; the elasticity of the lens is then diminished, and its dioptric properties change.

Chemical Composition

Krause has carefully analyzed the chemical composition of the lens and has determined the content of water, salts, protein, and fat (Table 103).

Table 103. Chemical Constituents of Crystalline Len

Constituent	*Per Cent*
Water	60–70
Ash	0.5–1
Lipids	3–6
Protein	35

In mammals, including man, the lens has 60 to 70 per cent of water; in fishes, only 50 to 60 per cent. With age this percentage diminishes and is less in the nucleus (60 per cent) than in the periphery (75 per cent), which is considered to be the youngest part of the lens. In the process of cataract formation, water at first increases and then diminishes when the cataract is "mature."

The ash of the lens represents 0.5 to 1 per cent of the total weight and has a composition similar to that of other tissues. There are relatively large amounts of potassium, corresponding to the large amount of intracellular water, and sulfur from cystine and glutathione. There are also Na, Cl, and small amounts of Mg and calcium. The calcium is a factor in its permeability and increases considerably in cases of cataract.

Berzelius (1825–1831) mentioned the existence of fats in the lens. In man the fat content increases up to 6 per cent at the age of twenty years and then remains stationary. Phospholipids form 80 per cent of the total fat; they are found together with protein mainly in the capsule. Cholesterol crystals are also found; they can be identified because they are doubly refracting.

Specific proteins, called "crystallins" by Berzelius, form 35 per cent of the weight of the lens. No other tissue has such a high protein content. There are six types of protein (Table 104). The amino-acid content of these proteins has not been accurately determined, but it is known that they do not contain glycin.

Table 104. Proteins in the Crystalline Lens of One-year-old Oxen

Protein	*Per Cent*	*Reported by*
Albuminoid	12.50	Mörner (1894)
α-Crystallin	31.74	Mörner (1894)
β-Crystallin	53.39	Mörner (1894)
Albumin	1.46	Mörner (1894)
Nucleoprotein	0.07	Krause (1933)
Mucoprotein	0.84	Krause (1933)

Source: Krause.

These proteins undergo *in vitro* and *in vivo* a process of autolysis which consists in disintegration with the release of a large quantity of amines (NH_2). Autolysis is produced by acids and enzymes. Kraus has extracted two proteases (α and β) from the normal lens, which act in acid medium. They are important in the genesis of cataract.

Proteins in the lens have peculiar antigenic properties. When injected they provoke the formation of antibodies specific for the lenses of all mammalian species (Uhlenhuth, 1903). The proteins contained in the lens of amphibians and birds differ in this respect from those of mammals. The so-called "α-crystallins" have a high antigenic power and the "β-crystallins" a low one.

The maintenance of normal structure in the lens is conditioned by dietary factors; thus tryptophane-free diets provoke in young rats the formation of cataracts and vascularization in the cornea.[1]

Physical Properties

The osmotic pressure of the lens is similar to that of blood. The freezing point varies from −0.51 to −0.63°C. Submerged in 1 per cent NaCl solution, the lens does not change in volume. The pH of the lens is 7.4, the same as that of other tissues.

The capsule of the lens is permeable to crystalloids but not to colloids. Glucose, lithium

[1] Schaeffer, A., and J. Murray, *Arch. Ophth.*, **43**, 202 and 1056, 1950.

salts, and even alcohol injected subcutaneously or given by mouth are found in the lens a few hours later. If the lens is submerged in water or a hypotonic fluid, it swells in proportion to the degree of hypotonicity of the fluid. Calcium salts diminish edema thus produced.

Light rays are partially absorbed as they pass through the lens. Approximately 8 per cent of the visible spectrum is absorbed, but not all wavelengths are absorbed equally. Infrared and ultraviolet rays are absorbed in a considerable proportion and can produce damage to the lens. The proteins of the lens are responsible for this absorption.

Physiology of the Lens

The lens plays an important part in accommodation, which will be dealt with further on. It has an active metabolism, which has been the object of a considerable amount of work.

Respiration. The lens takes up oxygen from the aqueous humor, which absorbs it through the cornea and from the ciliary bodies. If the lens is removed, oxygen pressure in the aqueous humor increases. The oxygen consumption of the lens of the dog *in vitro* is 0.2 to 0.5 cc. per min., approximately one-tenth that of resting muscle.

Ascorbic acid and glutathione seem to play an important part in the respiratory processes of the lens. There are considerable amounts of both these substances. The concentration of glutathione in the lens of the sheep has been found to be 343 mg. per 100 gm. of fresh substance; therefore much higher than in other tissues. It is found in the oxidized and reduced forms. The concentration of both these substances is greater in the peripheral part of the lens than in the central part; it diminishes with age. The lens forms glutathione, but not ascorbic acid, which comes from the ciliary body.[1]

Glutathione takes up oxygen and passes it to β-crystallin, both substances forming an oxidation-reduction system. If all the glutathione is removed from the lens, the latter no longer takes up oxygen, but respiratory processes again take place if glutathione is added and at a rate proportional to its concentration.

Ascorbic acid is taken up from the aqueous humor in its oxidized form and is returned in the reduced form, oxygen remaining in the lens. The concentration of ascorbic acid in the lens is higher than that of the aqueous humor, which in turn has a higher concentration of ascorbic acid than the blood. Synthesis of this substance is supposed by some workers to take place in the lens. Glutathione and ascorbic acid diminish in a lens with cataract.

[1] Kinsey, E. V., *Arch. Ophth.*, **46**, 441, 1951.

Intermediary metabolism. Energy is set free in the lens by the oxidation of glucose, which diffuses from the blood into the aqueous humor and is taken up by the lens, where it is found in a concentration of 60 mg. per 100 gm. After removal of the lens, the concentration of glucose in the aqueous humor increases up to the level of the blood sugar. The low content of glucose usually found in the aqueous humor is therefore due to sugar consumption by the lens. The concentration of glucose in the aqueous humor and the lens increases in diabetes. Approximately one-quarter of the total amount of sugar metabolized by the lens is completely oxidized to CO_2 and water; three-quarters of it is converted into lactic acid. There is a cycle similar to that found in muscle. Lactic acid formed accumulates in the lens and diffuses out into the aqueous humor, where it reaches a higher level than that of the blood. After removal of the lens, the concentration of lactic acid in the aqueous humor falls to the same level as in blood. Flavoprotein and the cytochrome–cytochrome oxidase system have been found in the lens (see Chap. 34, Cell Respiration).[1]

ACCOMMODATION OF THE EYE FOR OBJECTS AT DIFFERENT DISTANCES

A clear vision of objects placed at different distances from the observer is obtained by a series of changes in the eye, known as accommodation. The crystalline lens plays a fundamental part in this process.

The Need for Accommodation. Scheiner's Experiment

When an object that is focused on the plate or film of a photographic camera is moved nearer to, or farther from, the camera, the distance between the lens and the plate must be changed, because the focus is displaced in the same direction as the object. When an object is brought nearer to the eye, the light rays coming from the object are progressively more diver-

[1] Kinsey, E.V., and C. Frohman, *Arch. Ophth.*, **46**, 536, 1951.

gent, and they would be focused at a point behind the retina if the eye remained unchanged. The displacement of the focus is insignificant when the object is moved from infinity to a distance of 6 m. (20 ft.) from the eye, but at nearer distances and up to 12 cm. from the eye, changes (*i.e.*, accommodation) take place which focus the object clearly on the retina (Table 105).

Table 105. Distance of Object from Eye and Image from Cornea

Distance of Object from Eye, M.	*Distance of Image from Cornea, Mm.*
Infinity	20
5	20.06
1	20.30
0.50	20.62
0.25	21.27
0.12	23.57

Source: Luciani.

Scheiner in 1614 demonstrated accommodation by means of the following experiment: A needle is held at arm's length and looked at through two pinholes made in a black card distant 2 to 3 mm. from each other. The needle will be seen double if the eye is focused nearer or farther away, because the retina is in front of the focus of the needle in the first case and behind the focus in the second (Fig. 428).

Fig. 428. Illustration of Scheiner's experiment.

The Importance of the Crystalline Lens

Demonstration of the role of the lens in accommodation. Indirect proof of the role of the lens has been obtained by showing that the cornea and retina are not displaced in the process of accommodation; therefore the lens must be modified in order to change the focus.

Young (1801) pointed out that individuals who had no crystalline lens had no power of accommodation and had to use lenses of 3 D. to see clearly objects placed near the eyes. Direct proof of the changes that occur in the lens in the process of accommodation was given by observation of the Purkinje images by Langenbeck (1849), Cramer (1851), and especially Helmholtz (1851–1853). If a subject in a dark room has a lighted candle placed on one side at approximately 50 cm. from the eye, an observer placed on the other side, so as to see the light reflected by the subject's eye, will see three images of the candle:

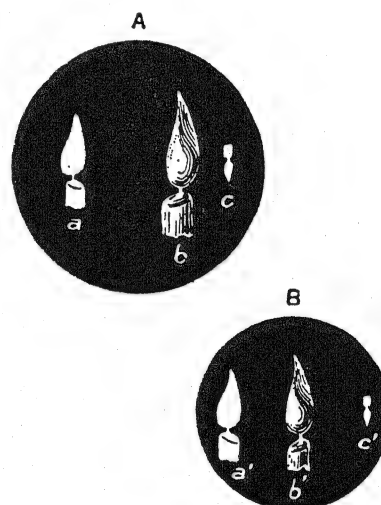



Fig. 429. Purkinje images. *A*, eye at rest; *B*, eye accommodated for near objects; *a* and *a'*, images reflected on the cornea; *b* and *b'*, images reflected on the anterior surface of the lens; *c* and *c'*, images reflected on the posterior surface of the lens.

1. The brightest image is reflected by the anterior surface of the cornea and is upright.

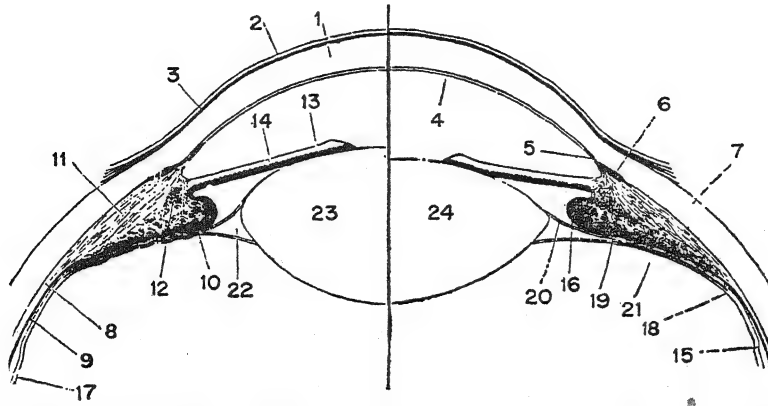



FIG. 430. Mechanism of accommodation. Left half, eye accommodated for near objects; right half, eye during vision of distant objects. 1, substantia propria of the cornea; 2 and 3, epithelial layer on the anterior surface of the cornea; 4, epithelial layer on the posterior surface of the cornea; 5, pectineal ligament; 6, spaces of Fontana; 7, sclera; 8, chorioid; 9, retina; 10, ciliary body; 11, ciliary muscle; 12, circular fibers of the ciliary muscle; 13 and 14, iris; 15, ora serrata; 16, ciliary processes; 17, hyaloid membrane; 18, zonula; 19 and 20, anterior and posterior surface of the zonula; 22, intra-zonular space; 23, lens accommodated for near vision (increased curvature of the anterior surface); 24, lens accommodated for distant vision.

2. The second image, also upright, is much dimmer but of larger size. It is reflected by the anterior surface of the lens.
3. The third image is the smallest of the three. It is bright and inverted, and due to reflection on the posterior surface of the lens, which acts like a concave mirror.

If the subject looks at an object near the eye, the first and third images do not change, the second image diminishes in size and approaches nearer to the first, owing to an increase in the curvature of the anterior surface of the lens (Fig. 429). This increase in the curvature of the lens pushes the iris forward toward the cornea, the anterior chamber of the eye diminishes, and the angle formed by the iris and the cornea becomes more acute.

Changes in the lens in accommodation. Changes in the Purkinje images have been accurately observed by means of the ocular microscope and registered photographically. If a near object is looked at, the curvature of the anterior surface increases in the middle part of the lens, but diminishes in the peripheral parts; the posterior surface also becomes slightly more convex. The diameter of the lens diminishes, and its anteroposterior axis increases (Fig. 430); the dioptric power increases by 12 D.

The speed with which these changes occur is greater when the eye is focused from a distance to a near object (0.39 to 0.82 sec.) than when accommodation takes place in the opposite sense (0.5 to 1.16 sec.).

The range of accommodation is measured in diopters, which give the refractive power of the eye. It decreases with age, especially after forty years (Fig. 431), when reading or any work that

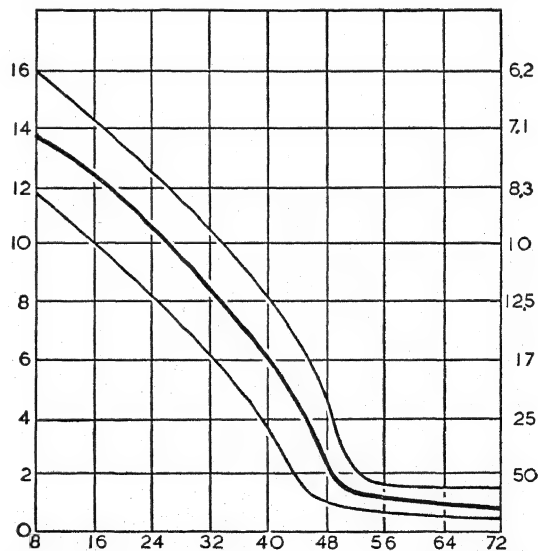



FIG. 431. Amplitude of monocular accommodation in man at different ages. Average, thick line; extremes, thin lines. Abscissa, age in years; left ordinate, diopters; right ordinate, distance of near point in centimeters. (*Duane, 1922.*)

requires near sight becomes difficult. This defect is known as presbyopia.

Associated phenomena. The pupil is constricted for near vision; thus marginal rays are eliminated and aberration is corrected. The increase in the depth of the focus also improves sight. The visual axes converge so that the images are formed on corresponding points of the retina.

Accommodation and convergence are closely associated and strengthen each other, but they can be artificially separated and are dissociated in certain diseases. Pupillary constriction is dependent on accommodation and convergence, but more on the latter than the former.

Action of the Zonula

Anatomy. The zonula is a fibrous ligament that extends from the ciliary body to the margin of the lens, being inserted on its anterior and posterior surfaces near the margin. Between the fibers there is jellylike substance which is easily permeable to fluids in both directions. The zonula separates the posterior chamber from the vitreous humor.

Physiologic action. The changes in shape of the lens are dependent in great part on the action of the zonula. In cases of congenital absence of the iris, the zonula can be observed with the ocular microscope, and it will be seen to relax on looking at a near object or when eserine is applied. The lens can then move slightly (0.3 to 2 mm.) when the subject's head is moved.

Action of the Ciliary Muscle

Anatomy. A transverse section of the eye shows the ciliary muscle as a triangle with its base on the iris and its apex in the chorioid at the level of the ora serrata. In man it is formed by two sets of smooth fibers. In one set the fibers are disposed in a radial or meridional direction; in the other, lying central to the first, the fibers follow a circular course. Each set has its own innervation.

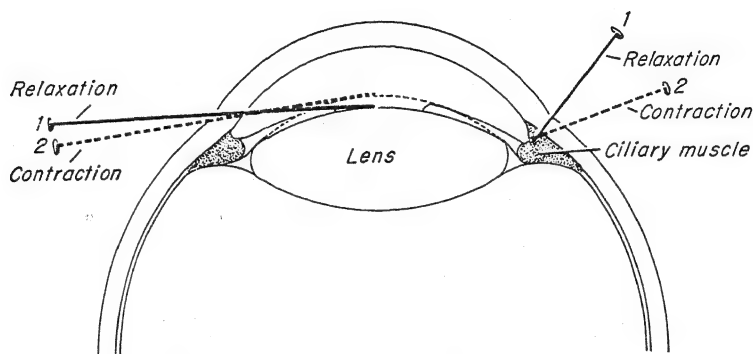



Fig. 432. Hansen and Volkers' experiments.

Physiologic action. The action of the ciliary muscle was first demonstrated by Hansen and Volkers in 1868; recently their work has been confirmed and developed by Sachs (1942). The muscle has been observed in dogs, cats, monkeys, and man directly through an opening made in the sclera for that purpose. On stimulation of the ciliary ganglion, the muscle will be seen to move approximately 0.5 mm. forward toward the cornea. This movement is also demonstrated by passing a pin through the sclera so that its point touches the muscle; when the latter contracts the head of the pin is displaced backward. If another pin is placed so that it passes over the anterior surface of the iris to the anterior surface of the lens, its head moves in the same direction as the first (Fig. 432). These experiments show that contraction of the ciliary muscle approximates it to the cornea and coincides with an increase in the convexity of the lens.

Innervation. Helmholtz maintained that accommodation to near objects was an active process, and accommodation to distant objects a passive one. There are, however, several proofs that both are active processes dependent on autonomic innervation. Olmsted's[1] experiments are conclusive in this respect. Stimulation of one of the divisions of the autonomic nervous system, after section of the fibers corresponding to the other division, in rats, dogs, rabbits, monkeys,

[1] Olmsted, J. M. D., *J. Nerv. & Ment. Dis.*, **99**, 794, 1944.

etc., has shown that the cervical sympathetic (through the long ciliary nerves) has a tonic effect, producing hypermetropia by flattening the anterior surface of the lens, thus accommodating the eye for distant vision. The parasympathetic (through the oculomotor and the short ciliary nerves) increases the convexity of the lens, producing myopia, thus accommodating the eye for near vision. The two divisions of the autonomic are reciprocal antagonists, but the effect of the parasympathetic is greater, being equivalent to 10 D., than that of the sympathetic, which is equivalent to 1.5 D.

Pharmacology. Drugs that act on visceral nerves modify the process of accommodation. Parasympathomimetic drugs, such as eserine, muscarine, and pilocarpine, produce accommodation for near vision, the intensity of their effect being in the order given. Parasympatholytic drugs, such as atropine and its derivatives, homatropine, euphthalmine, etc., paralyze accommodation. Both types of substances are often used by ophthalmologists.

Theories of Accommodation

Several theories have been proposed to explain accommodation. The Young-Helmholtz theory is widely accepted. According to this theory, when the eye is resting, the zonula is under tension and maintains the elastic capsule of the lens distended. When the eye is accommodated for near vision, the following changes occur: (*a*) the ciliary muscle contracts and pulls the chorioid forward; (*b*) the zonula is therefore relaxed; (*c*) the lens, owing to the elasticity of its capsule, takes on a more rounded shape. All these facts have been satisfactorily demonstrated, but the theory does not explain certain other facts; *e.g.*, the fact that the anterior surface of the lens takes on a hyperbolic and not a rounded shape. Moreover, Fincham has seen that the capsule is not of uniform thickness over all the lens, but is thinner over the central part of the anterior surface, so that the curvature can increase more at this level than in the periphery.

INTRAOCULAR FLUIDS AND INTRAOCULAR PRESSURE

The intraocular fluids, *i.e.*, the aqueous and vitreous humors, are of similar chemical composition, and both are contained in the eyeball under pressure.

Aqueous Humor

Aqueous humor is a colorless fluid, similar to water, which circulates in the anterior and posterior chambers of the eye. It is a nutrient fluid for the lens and the vitreous humor, and it serves to regulate intraocular pressure.

The amount of aqueous humor in different species and its relation to the volume of the eyeball are given in Table 106. In man normally there is approximately 150 cu. mm., and in myopic eyes there is 250 cu. mm. Its chemical composition is given in Table 107. It has much less protein and more water than the blood plasma. Other substances are found in more or less the same concentration in the aqueous humor as in blood plasma, except that there are less glucose, amino acids, and urea, and more salts, especially chlorides. There is also oxygen and CO_2 in solution.

Table 106. Amount of Aqueous and Vitreous Humors of the Eye in Different Species

Species	*Aqueous humor, cu. mm.* (*Magitot,* 1921)	*Vitreous humor, cu. mm.* (*Emmert,* 1896)	*Anterior chamber, % volume of eyeball* (*Wessely,* 1921)
Man	150	......	4
Rabbit	270–300	2,000	12
Cat	340–350	2,800	22
Dog	400–500	3,200	12.5
Calf	800–850	11,820	

Table 107. Composition of Aqueous and Vitreous Humors and Blood Serum

Constituent	*Aqueous humor, gm. per* 100 *cc.*	*Vitreous humor, gm. per* 100 *cc.*	*Blood serum, gm. per* 100 *cc.*
Water	99.692	99.682	93.323
Solids	1.086	1.108	9.536
Protein	0.020	0.065	7.369
Fat	0.004	0.007	0.130
Glucose	0.098	0.098	0.091
Sodium	0.278	0.273	0.335
Chloride	0.437	0.416	0.366
Potassium	0.018	0.019	0.020

Source: Duke-Elder.

Formation of aqueous humor. The aqueous humor is formed from blood plasma in the same

way as other tissue fluids. Thus diffusible substances injected into the blood (fluorescein, potassium ferricyanide, etc.) or taken by mouth (potassium iodide) are found soon after in the auqeous fluid. The process of formation is still a subject of discussion.

Aqueous humor is formed on the surface of the iris, the ciliary body, and the cornea, according to the work of Kinsey and Cogan previously discussed. There is now definite proof that all these structures contribute to the formation of the aqueous humor.

An experiment performed by Ehrlich in 1881 showed that aqueous humor is formed on the surface of the iris. Fluorescein (2 to 3 cc. of a 20 per cent solution) is injected intravenously into a rabbit, and 2 or 3 min. later a greenish cloud is seen to form on the anterior surface of the iris. In some cases aqueous cysts are formed on this surface. The iris is not the only site of formation of the aqueous humor, because this fluid continues to be produced after extirpation of the iris and in cases of congenital aniridia (absence of the iris).

Aqueous humor is also formed on the surface of the ciliary body. If the pupil is closed, experimentally or by disease, aqueous humor accumulates in the posterior chamber, which is again filled after it has been evacuated by puncture. Aqueous humor continues to be formed in cases of congenital absence of the ciliary body; therefore it is not formed exclusively on these structures.

The mechanism by which the aqueous humor is formed (secretion by the ciliary body or filtration or dialysis) has been the object of much discussion. According to Duke-Elder,[1] some substances are filtrated and others are dialyzed.

Intraocular circulation. Aqueous humor circulates through the chambers of the eye and is evacuated by three routes: (*a*) the posterior route; (*b*) the iris; (*c*) the canal of Schlemm. The posterior route through the zonula, the canal of Cloquet (which extends from the posterior surface of the lens to the disk), and the sheath of the optic nerve accounts for only 1 per cent of the total fluid. The crypts of the anterior surface of the iris near its blood vessels also reabsorb part of the aqueous humor.

The canal of Schlemm is a venous sinus which surrounds the cornea like a ring at the level of the angle formed by the iris and the cornea. It is separated from the anterior chamber by trabeculae which come from the cornea and form the spaces of Fontana, which communicate with the anterior chamber but not with the canal of Schlemm. Small particles, such as those of India ink, injected into the anterior chamber accumulate in the iridocorneal angle, also known as the filtration angle. According to some observers, the canal of Schlemm is continuously reabsorbing aqueous humor, but according to others it is a safety mechanism that functions only when the intraocular pressure increases, the eyelids contract, etc.

The circulation of the aqueous humor is carried out at a relatively fast rate. This has been demonstrated by the injection of "labeled" substances (*e.g.*, isotopes) which can be followed in their passage through the eye. The aqueous humor has thus been seen to be completely renewed several times per day. Kinsey and Cogan maintain that all the water is renewed every 3 min. In the cat and rabbit renewal of the aqueous humor takes place at a rate of 1 to 2 per cent per min.[1] Stains can also be seen to disappear rapidly, but many factors other than the circulatory rate (*e.g.*, phagocytosis) condition the reabsorption of these substances. If the anterior chamber is evacuated, it is refilled in 40 to 50 min. by a fluid that has a higher protein concentration and less chloride than the normal fluid.

There is also a thermal current of fluid, which flows up the anterior surface of the iris and down the posterior surface of the cornea. It is due to the difference in temperature between the iris and the cornea; the latter has a temperature 7°C. below that of the former. This current is clearly seen after the injection of fluorescein.

Vitreous Humor

The vitreous humor is a colorless, transparent, jellylike fluid which fills the space between the posterior surface of the lens and the retina. It contains a network of thin fibers, which can be seen *in vivo* with the ocular microscope. They form the hyaloid membrane on the surface of the retina. If the vitreous humor is centrifuged, the fibers fall to the bottom and a fluid identical to aqueous humor is separated.

[1] Duke-Elder, W. S., Nature of the Aqueous Humor, *J. A. M. A.*, **137**, 1285, 1948.

[1] Langham, M., *Brit. J. Ophth.*, **35**, 409, 1951; Ross, E., *Brit. J. Ophth.*, **36**, 41, 1952.

The density of vitreous humor (1.005 to 1.008) and its freezing point are similar to those of the aqueous humor. Its chemical composition (Table 107) differs from that of aqueous humor by its higher protein content. Two proteins have been separated, a mucoprotein (mucoid) and vitrein, similar to gelatin, which is highly hygroscopic and forms fibers. These proteins give the vitreous humor the consistency of a gel, which is denser near the retina and less dense on the posterior surface of the lens. Hyaluronic acid is found in relatively large amounts. This mucopolysaccharide probably forms part of the mucoid. It absorbs great quantities of water and contributes to give viscosity to the fluid. Hyaluronidase, which is found in the ciliary body, iris, cornea, etc., disintegrates hyaluronic acid and fluidifies the substances which contain it. Piris injected this enzyme in the vitreous humor of living rabbits and observed liquefaction of the vitreous humor and a fall in intraocular pressure which lasted for several days.

Intraocular Pressure

The fluids in the eye are contained in a cavity formed by the inelastic sclera. They are under pressure, which results from an equilibrium between the inflow and outflow. The pressure falls rapidly after death or if the eye is separated from the body.

The normal intraocular pressure and its physiologic variations. Intraocular pressure can be measured directly by introducing a needle into the eyeball and connecting it with a manometer. It can also be determined by digital pressure or by means of tonometers. The latter are instruments that measure the pressure needed to flatten the cornea or sclera. The normal pressure in the human eye, measured by a tonometer, is usually 20 to 25 mm. Hg, the extreme values being 12 and 35 mm. Hg.

Intraocular pressure diminishes gradually throughout life, from childhood to old age. It is 3 to 5 mm. higher in the morning than in the evening. Respiratory movements and the arterial pulse cause slight rhythmical oscillations in intraocular pressure. When the lids are closed, the pressure rises approximately 10 mm. Hg.

Causes of intraocular pressure. Intraocular pressure is due to the pressure in the blood capillaries of the chorioid membrane, which depends on the general arterial blood pressure.

Arterial blood pressure. Intraocular pressure is related to the arterial blood pressure and rises and falls with it. The relation between intraocular pressure and arterial blood pressure in different species is given in Table 108. If the blood pressure in the chorioid capillaries falls, *e.g.*, owing to clamping of the carotids, hemorrhage, bradycardia (stimulation of the vagus nerve), removal of the eyeball from the body, or death, intraocular pressure also falls. If the blood pressure rises, *e.g.*, owing to clamping of the abdominal aorta, or intravenous injection of a large quantity of fluid, intraocular pressure also rises.

Table 108. Intraocular Pressure and Blood Pressure in Different Species

Species	*Intraocular pressure, mm. Hg*	*Blood pressure, mm. Hg*
Guinea pig	10–15	110–120
Rabbit	18–25	100–140
Cat	23–30	130–150
Dog	15–25	140–180
Man	15–30	110–120*

* In femoral artery.

Circulation in the chorioid. Variations in the pressure and flow in the chorioid vessels have a direct influence on intraocular pressure. The circulation in the eye is divided into two systems: (*a*) the retinal vessels, which are branches of the central artery of the optic nerve; (*b*) the chorioid vessels, which are branches of the ciliary arteries. The latter is the only system of importance for intraocular pressure. The chorioid is a vascular membrane situated between the almost rigid sclera and the retina. Its thickness is approximately 300 μ. The ciliary arteries enter this membrane and immediately branch out, forming a rich network of large capillaries (their diameter is approximately 10 times that of other capillaries in the organism). The pressure in the ciliary arterioles is approximately 80 mm. Hg; owing to the large size of the capillaries, a relatively great part of this pressure is transmitted to the veins, where the blood pressure is approximately 26 mm. Hg. The pressure in the capillaries lies between these two figures. This pressure is maintained by the rigid sclera; therefore any increase in the pressure within the capillaries causes an increase in intraocular pressure,

and vice versa, a fall in capillary pressure is followed by a fall in intraocular pressure. Ligature of the ciliary arteries provokes a fall in the pressure of the capillaries in the chorioid and in intraocular pressure. On the other hand, ligature of the veins increases capillary and intraocular pressure.

Variations of intraocular pressure with respect to variations in the diameter of the chorioid capillaries are related to the changes in capillary pressure. Thus if the capillaries are dilated by subconjunctival injection of hypertonic NaCl solution, intraocular pressure rises slightly, but if later pituitrin is injected so that a considerable rise in arterial blood pressure occurs, the intraocular pressure also rises considerably. On the other hand, if the chorioid capillaries are constricted by adrenaline, intraocular pressure does not rise even if the general blood pressure rises. Changes in the diameter of the capillaries therefore influence intraocular pressure in so far as they modify capillary pressure.

Effects of innervation. Nerve impulses modify intraocular pressure only through the changes they provoke in capillary pressure. Stimulation of the cervical sympathetic causes vasoconstriction, and the intraocular pressure falls considerably. Section of the cervical sympathetic causes a slight transitory rise in intraocular pressure

Stimulation of the trigeminal nerve raises intraocular pressure, and section of this nerve lowers it. Axon reflexes take place through the endings of this nerve. After it has been cut, stimulation of the iris with a needle (Magitot) or subconjunctival injection of hypertonic solution provokes an increase in intraocular pressure, which is annulled by the injection of adrenaline, causing vasoconstriction.

Drugs that act on autonomic effectors also cause changes in intraocular pressure by modifying the pressure in the chorioid capillaries. Thus the following substances instilled into the conjunctiva modify intraocular pressure: (*a*) adrenaline diminishes it after a slight transitory increase; (*b*) eserine increases it; (*c*) atropine has no effect on it in normal subjects, but increases it in patients with glaucoma[1] and should not be used in these cases.

BIBLIOGRAPHY

See page 976.

[1] Glaucoma is a disease of the eye in which there is a high intraocular pressure, which disturbs the nutrition of the eye and usually ends in blindness.

CHAPTER 78

The Retina. The Visual Pathways and Centers

THE STUDY OF THE embryology and histology of the retina shows that it is a specialized part of the central nervous system.

THE RETINA

In recent years the classic work of Cajal on the histology of the retina has been considerably developed; the observations of Polyak are particularly important in this respect. The photochemical processes and electrical phenomena that take place in the retina have been studied intensively and have thrown much light on the physiology of the organ of vision.

Anatomy and histology. Histological examination of the retina shows that it is formed by two concentric parts: (*a*) under the chorioid membrane there is a pigmented epithelium; (*b*) below this epithelium, in close contact with the vitreous humor, there is the pars optica. The pigmented epithelium is formed by a layer of pigmented cells, which convert the interior of the eyeball into a camera obscura.

The pars optica is the sensory part of the retina, and it is transparent in the living eye. Its thickness is 0.6 mm. at the back of the eye and diminishes toward the anterior part, so that it is only 0.1 mm. in the ora serrata, where it ends. It is perforated by the entrance of the optic nerve. Slightly above (1 mm.) and 3 mm. toward the nasal side of this entrance, on the visual axis of the eye, there is a small yellowish area (approximately 3 sq. mm.), called the macula lutea. In the center of the macula there is a depression, the fovea centralis, which is the most sensitive part of the retina for daylight vision. The pars optica is formed by nine layers, in which there are three fundamental types of cell: (*a*) photoreceptors; (*b*) association cells; (*c*) ganglion cells (Fig. 433).

The photoreceptors are in the outermost layer of the pars optica in contact with the pigment layer and placed vertically with respect to this membrane. There are two kinds of photoreceptors, the cones and the rods, which nevertheless have a common embryonic origin. These cells are formed by three parts, an external photosensitive process, similar to a cilium; a middle part, which contains the nucleus; and an internal process, which is connected with the association cells (Fig. 434). There are approximately 7,000,000 cones in the human retina. They are the only type of receptor in the fovea, where their length is 80 μ. Toward the ora serrata they diminish in size and number, and are surrounded by rows of rods. The external process is short except in the fovea, where it is almost as long as that of the rods. There are approximately 130,000,000 rods in the human retina; they increase in number toward the ora serrata. They have a long, fine, external process which contains 10 per cent rhodopsin. The fine internal process is connected with association cells.

There are three types of association cells distributed in the middle layers of the pars optica. The bipolar cells have two processes; the external process is connected with the internal process of the photoreceptors, and the internal process ends on the ganglion cells. Other association cells are the horizontal and amac rinal cells, which are also connected with the photoreceptors and the ganglion cells.

The ganglion cells are placed in the innermost layer of the retina. They have a short process connected with the bipolar cells and a long process which is a naked axon forming the optic nerve.

The cells in the retina, therefore, form a chain with three links, *i.e.*, the photoreceptor (cone or rod), the bipolar cell, and the ganglion cell. Most of the bipolar and ganglion cells have multiple connections whereby

they receive impulses from several, in some cases many, photoreceptors (Polyak's "polysynaptic" cells). A few of the bipolar cells are much smaller (midget, or monosynaptic, bipolar cells); they are found in the fovea centralis and are connected with only one cone and one ganglion cell. In the fovea there are also small ganglion cells which receive impulses from only one or two cones by way of the midget bipolar cells (Fig. 433).

Fig. 433. Principal connections of the cells in the retina.

Structural Changes Produced by Light

In fishes, amphibians, reptiles, and birds, light provokes in the retina photomechanical changes which involve the pigment cells and the photoreceptors. If a frog or toad is exposed to a bright light, the pigment granules in the pigment cells are rapidly dispersed so that within a few minutes they surround the external processes of the photoreceptors. When the animals are placed in a dark environment, the pigment granules are slowly concentrated around the nucleus of the pigment cells (Fig. 435).

Dispersion and concentration of the melanosomes (pigment granules) are controlled by a complex neuroendocrine mechanism in which the pars intermedia of the hypophysis plays an important part (see Chap. 52).

In fishes and amphibians, and to a lesser extent in reptiles and birds, the rods are shortened when the animals are exposed to light. This process is even more marked in the cones, which are sometimes shortened to one-tenth their resting length. This shortening is due to the contraction of the middle part of the cell, which has a myoid structure. Apparently the shortening serves to focus the light on the photoreceptor. In these species the external processes of the rods are dilated on exposure to light, and they may come into contact with the neighboring rods.

Chemical Changes Produced by Light

The discovery of photography led to the belief that a similar process took place in the eye. The change in the color of the retina produced by light, which was reported a few years later (1877) by Boll, seemed to confirm the idea. Kühne (1879) then discovered rhodopsin and carefully investigated the part it played in visual processes. Recent advances in the physics and chemistry of photosensitive substances have given new and more accurate knowledge on this subject.

Origin of visual pigments. Photosensitive substances, which play a part in the visual processes of animals and in phototropism in plants, form part of the group called carotenoid pigments; one of these,

β-carotene, is closely related to vitamin A (see Chap. 49).

Carotenoids vary in color from yellow to red. Phototropic plants have large amounts of carotenoid pigments in the parts sensitive to light. There are no carotenoid pigments in plants that are not phototropic. These pigments are sensitive to the blue radiations of the spectrum, while the chlorophyll system is sensitive to yellow-orange radiations.

Wald[1] has drawn attention to the fact that these substances are utilized by plants and animals (in phototropism and visual processes, respectively) but only plant organisms synthesize them. Animals cannot synthesize carotenoids; they obtain them from plants and utilize them after only slightly modifying their structure. Animals are, therefore, dependent on plants for the supply of these pigments.

FIG. 434. Photoreceptors in the retina. Diagram of a rod (left) and a cone (right).

There are several visual pigments in the cones, the rods, and the macula. The best known are the pigments of the rods, rhodopsin and porphyropsin.

Rhodopsin. This red pigment, also called "erythropsin" or "visual purple," is bleached by exposure to light and recovers its color in the dark. It is found in the external segments of the rods.

Rhodopsin plays a fundamental part in nocturnal or twilight vision,[2] which is a function of the rods. The different radiations of the spectrum have the same effect on the dark-adapted eye as on visual purple. Thus the sensibility curve of the dark-adapted eye coincides with the absorption curve of a pure solution of visual purple (Fig. 436).

Bleaching. When rhodopsin is exposed to light it is rapidly bleached, passing through a series of

[1] WALD, G., *Harvey Lect.*, **39**, 117, 1945–1946.

[2] Nocturnal or twilight vision is given by illumination such that objects are seen in outline or as shadows, without any distinction of color.

FIG. 435. Retina of normal toad [*Bufo arenarum* (Hensel)] kept in the light and in darkness. Dispersion of pigment in light (left) and concentration in darkness (right).

colors, transitory orange and visual yellow, until it becomes white. These changes are easily followed not only *in vitro* in solutions of rhodopsin, but also *in vivo*, by cutting out part of the sclera. The process can also be demonstrated by the following experiment: A rabbit is kept in the dark for several hours and then placed in front of a brightly lighted window for a few minutes. The animal is then killed and the eye submerged in alum. A white patch will be seen in the retina where the image of the window was formed, and if the window had bars that threw shadows, these will appear as dark lines on the white patch.

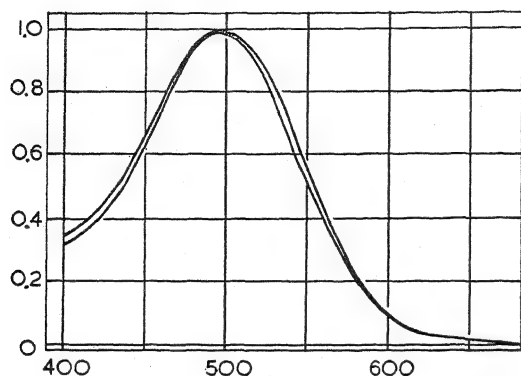



FIG. 436. Corrected curves of scotopic vision of the human eye and absorption curve of a solution of visual purple. Abscissa, wavelength in millimicrons; ordinates, luminosity and absorption. (*Hecht and Pirenne.*)

Rhodopsin can be extracted by bile salts from the retina of an animal which has been kept several hours in darkness.[1] It is made up of a large protein molecule (molecular weight 270,000), *opsin*, which is specific for each animal species. It has a carotenoid group, equivalent to one-tenth of the weight of the whole molecule, to which it owes its red color and photosensitiveness. Absorption spectra of rhodopsin show three bands, α, β, and γ at 500, 350, and 278 mμ. The α band is the highest; it is due, together with the β band, to the carotenoid. The γ band is due to the opsin.

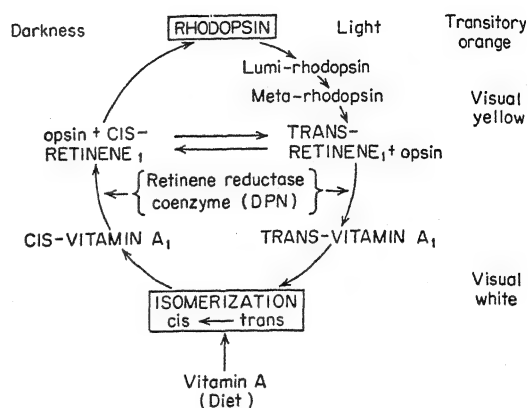



FIG. 437. Chemical cycle of rhodopsin.

Exposure to light causes profound physical and chemical changes in rhodopsin.[1] In the first stage the carotenoid is separated from the opsin (Fig. 437). Rhodopsin is rapidly converted to *lumi-rhodopsin*, with the absorption spectrum shifted 5 mμ toward the ultraviolet; then to *meta-rhodopsin*, with a further shift of 5 mμ in the spectrum, and finally it is split up into opsin and retinene$_1$, which has a characteristic absorption spectrum (Table 109). This substance is an aldehyde of vitamin A_1.[2] This stage ends with the formation of visual yellow.

Retinene$_1$ is converted by reduction into vitamin A_1. The enzyme system of this reaction is constituted by *retinene reductase* with DPN (diphosphopyridine nucleotide) as coenzyme. At the end of this stage visual white appears.

Table 109. Absorption Maxima of Photosensitive Substances in the Eye

Substance	*Maximum, mμ*	*Substance*	*Maximum, mμ*	*Substance*	*Maximum, mμ*
Pigment	467	*Porphyropsin*	522	*Iodopsin*	555
Rhodopsin	500	Retinene$_2$	405		
Retinene$_1$	387	Vitamin A_2	355		
Vitamin A_1	328				

Note: Rhodopsin and porphyropsin are dissolved in 1 per cent aqueous solution of digitonin. Retinenes and vitamin A are dissolved in chloroform.

[1] HECHT, S., The Chemistry of Visual Substances, *Ann. Rev. Biochem.*, **11**, 465, 1942.

[1] WALD, G., *Science*, **113**, 287, 1951; *Bull. Am. Acad. Arts & Sc.*, **6**, No. 3, 1952.

[2] MORTON, R. A., and T. W. GOODWIN, Preparation of Retinene in Vitro, *Nature, London*, **153**, 405, 1944.

Regeneration. Retinene$_1$ and vitamin A$_1$ produced by the decomposition of rhodopsin are mainly in the form of the *trans* isomer, which cannot be used in the regeneration of rhodopsin. The process of conversion of vitamin A$_1$ into the *cis* isomer is still unknown. This *cis* isomer of vitamin A$_1$, coming from the retinal cycle, together with some taken from the blood, is converted into retinene$_1$ by oxidation of the alcohol into aldehyde. The above-mentioned enzyme system activates this reaction. The *cis*-retinene$_1$ combines in darkness with the SH groups of opsin, and rhodopsin is regenerated.[1]

Subjects suffering from avitaminosis A$_1$ have a more or less marked deficiency in dark adaptation, which is quickly improved by the ingestion of vitamin A$_1$. Night blindness or a deficiency in twilight vision is a serious handicap in certain occupations (*e.g.*, aviators flying at night).

Other factors also condition regeneration of visual purple. Among these, oxygen partial pressure can be mentioned. When there is a low oxygen pressure, *e.g.*, when flying at great heights, regeneration is retarded and twilight vision is deficient.

Porphyropsin. In fresh-water fishes the rods contain porphyropsin instead of rhodopsin. This is a purple pigment which is bleached by light and is regenerated in darkness. It differs from rhodopsin in setting free retinene$_2$ instead of retinene$_1$ when it is bleached. Crystalline retinene$_2$, identical with that obtained from the retina, has been prepared from vitamin A$_2$ extracted from fresh-water fishes.[2] Given by mouth to rats suffering from avitaminosis A, it is rapidly stored in the liver; signs of deficiency decrease and the animals improve, but normalization is not complete. The part played by vitamin A$_1$ is taken on by vitamin A$_2$; otherwise the process is the same. There are slight differences in the chemical composition of vitamin A$_2$ and retinene$_2$ with respect to those of vitamin A$_1$ and retinene$_1$. There are also differences in the absorption spectra (Table 109).

Comparison of rhodopsin and porphyropsin. The two systems are very similar in their chemical aspects, but they are distributed among different species. Porphyropsin is found only in the eyes of fresh-water fishes, while rhodopsin is found in salt-water fishes and all other species. Wald has examined fishes that at certain stages of their life cycle are found in salt water and at others in fresh water (eels, salmon, etc.). He concludes that these animals have both systems but that the system corresponding to the place of ovulation is predominant. Wald has observed a transition from one system to the other in *Rana catesbiana,* which like other amphibians ovulates and completes its larval development in fresh water. Porphyropsin is found in the eyes of the larvae, rhodopsin appears in the course of metamorphosis, and in the adult form porphyropsin disappears and only rhodopsin is found.

[1] Wald, G., and P. Brown, *J. Gen. Physiol.*, **35**, 797, 1952.

[2] Cama, H., R. Morton, *et al.*, *Biochem. J.*, **52**, 535, 540, and 542, 1952.

Other visual pigments. Wald has extracted two pigments from the retina of the chicken, in which cones are predominant. One is probably rhodopsin; the other is called "iodopsin" because of its purple color. The chemical structure of this substance has not yet been definitely established, but it appears to be related to vitamin A and similar to rhodopsin. It has maximum absorption at 555 mμ, *i.e.*, the wavelength to which the light-adapted eye is most sensitive.

A pigment with maximum photosensitiveness at 467 mμ, and insensitive to red light, has been extracted from the retina of a fresh-water fish, the tench (*Tinca tinca*).[1]

In the macula lutea of men and monkeys a yellow pigment, distributed in all the layers of the retina, has been found. It acts as a filter absorbing blue and violet radiations, thus correcting chromatic aberration. This pigment is one of the xanthophyll group, similar to others in the group found in many plants and the outer covering of crustacea.

Electrical Activity of the Retina

The development of sensitive and accurate methods for registering electric potentials of very low voltage and short duration has advanced the knowledge of the functions of the retina, as in other nerve structures.

Resting potential. The eyeball *in situ* or separated from the organism shows a steady corneoretinal potential of 7 to 9 mv., the back of the eye being electronegative with respect to the cornea. The potential is apparently due to the retina, since it is suppressed by destruction of the retina and persists after degeneration of the optic nerve.

Action potentials. Electroretinogram. If the retina is stimulated by light, a phasic po-

[1] Dartnall, H., *J. Physiol.*, **116**, 257, 1952.

tential variation is observed which is known as the electroretinogram (ERG). Einthoven, who was the first to register this phenomenon, distinguished four waves, *a*, *b*, *c*, and *d* (Fig. 438).

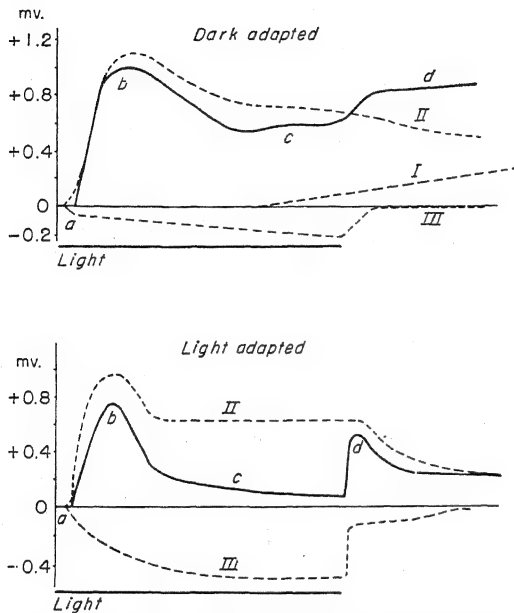



Fig. 438. Electroretinogram in light and darkness.

The first wave (*a*) is a small negative deflection which appears 0.01 to 0.05 sec. after the application of the light stimulus. The other three waves are positive deflections. *b* is a large positive wave followed by a second prolonged positive wave (*c*). When the stimulus ceases, a third positive wave (*d*) is seen on the slowly declining potential curve.

The *b* wave varies considerably. Its threshold coincides with that of scotopic vision. It increases with the intensity of the stimulus. It is large when the eye is stimulated after having been kept in the dark, and small when provoked by an increase in illumination. It is modified by several drugs.

The electrical activity of the retina begins when, in the course of embryonic development, the photoreceptors are formed. It is not dependent on the optic nerve.

In man the ERG can be recorded by applying the electrodes to the tissues surrounding the eyes. The *b* and *c* waves are well marked, but the *a* and *d* waves are not found. The latent period may be as long as 60 sec. (Bernhard).

Action potential of the optic nerve fibers. Hartline and his associates have registered the potentials of a single optic nerve fiber in the retina of the horseshoe crab (*Limulus*). They removed the cornea, lens, and vitreous humor and under the microscope separated one of the nerve fibers, which have no myelin sheath at this level. The microelectrode placed on the fiber was connected by means of amplifiers with a suitable recording system. When the retina was stimulated with white light, after a short latent period, a series of spikes of uniform amplitude appeared.

By stimulating limited areas of the retina, using a fine ray of light, it was possible to determine that each fiber is distributed over an area of approximately 1 sq. mm., which has hundreds of photoreceptors. There are three groups of fibers. When the stimulus is applied, one of the groups responds to light with a rapid discharge of spikes, which gradually declines to a steady state ("on" type of receptors). Another group responds to the cessation of light ("off" type), and a third group discharges at the beginning and end of stimulation ("on-off" type).

The influence of intensity and duration of the stimulus. The frequency and duration of the discharge increase with the intensity of the stimulus (Fig. 439). The high initial frequency gradually declines to a steady state (adaptation). If the intensity is kept constant, an increase in the duration of the stimulus provokes a more frequent discharge. The nerve centers cannot discriminate variations in the intensity and duration of the stimulus with the data obtained by stimulation of a single fiber; probably notation of these characteristics in a stimulus results from integration of discharges that travel along several fibers.

The influence of the wavelength of light. Granit has developed a method for the registration of potentials in a single fiber of the retina of higher animals. The anterior segment of the eyeball is removed, as in Hartline's experiment, or a window is made in the sclera. A microelectrode of platinum isolated by glass is placed on the fiber, which is isolated from the rest by means of a micromanipulator. The retina is then stimulated by a thin ray of light of known wavelength. The impulses are amplified and recorded or converted into sound.

Approximately 60 per cent of the fibers respond to the same wavelengths as the normal

eye. These are known as "dominators," and in the light-adapted eye the sensibility curve of these fibers is the same as that of the whole eye. If the eye undergoes dark adaptation, both sensitivity curves undergo a similar change, and the apex of sensitiveness passes from 555 mμ to

FIG. 439. Influence of duration and intensity of stimulus on the action potential of a single retinal fiber of the optic nerve of *Limulus*. (*Hartline.*)

500 mμ. Fibers that are dominators in photopic vision are also dominators in scotopic vision. These fibers are probably connected with cones

FIG. 440. Synthetic curve of photopic luminosity in man. (*Granit.*)

and rods in the retina and conduct impulses that give rise to the sensation of brightness or luminosity.

The rest of the fibers respond only to a limited wavelength. There are three main types, which are sensitive to red-yellow, green, and blue, respectively, *i.e.*, to the three primary colors. These fibers are called "modulators"; they are probably connected with cones and give rise to the sensation of color. The summation curve of sensibility of the three types of modulators is equal to the curve of photopic dominators (Fig. 440).

Correlation of chemical and electrical phenomena. Recently Ball and his associates[1] have observed that the sensibility curve of photopic dominators coincides with the absorption curve of iodopsin, and the curve of scotopic dominators with the absorption curve of rhodopsin. The curve of modulators coincides with the absorption curves of retinene$_1$ and vitamin A_1. This remarkable connection between chemical and electrical phenomena seems to indicate that light decomposes photosensitive substances in the retina, thus stimulating the receptors and discharging impulses along the optic nerve, which on arriving at the striate cortex give rise to visual sensations.

VISUAL PATHWAYS AND CENTERS

The visual pathway extends from the retina to the area striata of the occipital cortex. It is formed by a chain of three neurons with a subcortical center interposed.

[1] BALL, S., D. COLLINS, R. A. MORTON, and A. STUBBS, Chemistry of Visual Processes, *Nature*, 161, 424, 1948.

Reticulogeniculate Pathway

The *first-order neurons* in the visual pathway are the bipolar cells in the retina, which are stimulated by the photoreceptor (cone or rod) and transmit impulses to the *second-order neurons*, *i.e.*, the ganglion cells. The axons of the ganglion cells converge to the papilla or disk, which is the optic-nerve head. The papilla is the blind spot of the visual field, because there are no photoreceptors at this point. The fibers from the nasal side of the retina converge to the papilla following a straight course, like the spokes of a wheel. Fibers from the macula also go directly to the papilla, but as no fibers pass through the fovea, those from the temporal side of the retina form two loops, one above and the other below the fovea, separated by a sharp horizontal line which goes from the fovea to the temporal margin of the retina (Fig. 441).

In the optic nerve the fibers are also distributed systematically. The nasal retinal fibers are situated on the median or internal side of the nerve, and the temporal retinal fibers on the lateral or external side, separated into an upper and lower quadrant by the large bundle of macular fibers (Fig. 442).

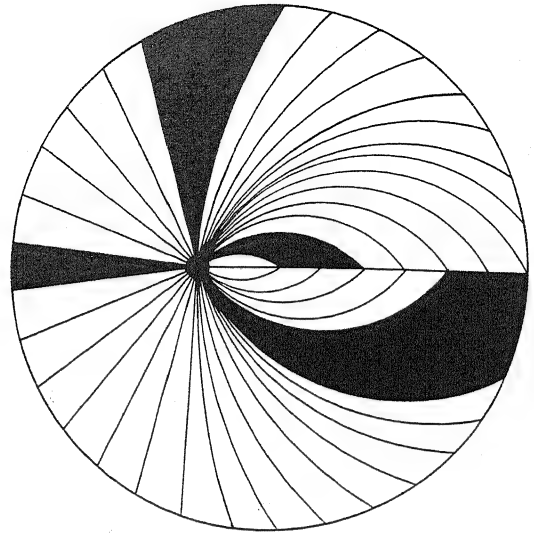



Fig. 441. Diagram of ganglion fibers in the retina. (*After Traquair.*)

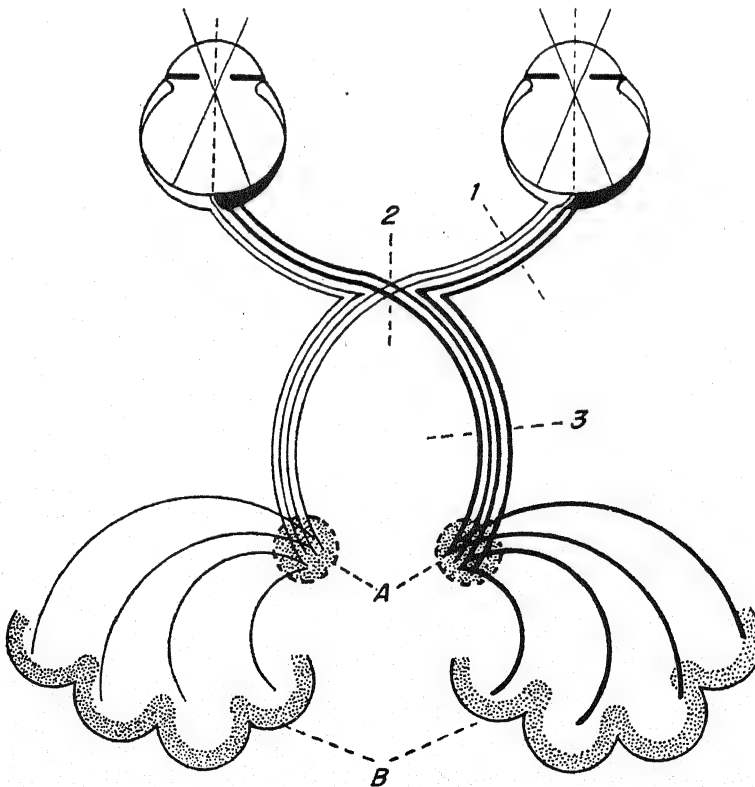



Fig. 442. Diagram of the visual pathway. *A*, lateral geniculate body; *B*, occipital cortex. The subcortical reflex paths have been omitted.

At the optic chiasm, fibers from the temporal hemiretina pass into the optic tract of the same side; those from the nasal hemiretina decussate and pass into the optic tract of the opposite side. Macular fibers are divided at the level of the chiasm into two bundles. Those from the temporal half pass into the homolateral optic tract, and those from the nasal side into the heterolateral optic tract.

The optic tract ends in the lateral geniculate body. In this tract a functional regrouping of the fibers occurs. Visual fibers, *i.e.*, those carrying impulses that provoke visual sensations, enter the lateral geniculate body, while the afferent fibers of the pupillary light reflexes end in the pretectal region. Some of the macular fibers are bilateral.

Geniculocortical Pathway

The lateral geniculate body is made up of six layers of cells, separated by fibers. The second-order neurons end in this body and are distributed systematically, so that the retina is projected on corresponding parts of the geniculate body. Thus macular fibers end in a large sector interposed between sectors with fibers from the lower quadrant of both retinas on the lateral side and those from the upper quadrant of both retinas on the medial side. Each optic-nerve fiber divides into several branches, and each branch ends on the body of a third-order neuron which receives no other fibers. These neurons are thus stimulated through a single synapse.

Fig. 443. Diagram of projection of retinal areas on the area striata.

A lesion of the retina or section of the optic nerve is followed by degeneration in the optic tract of the cut fibers, which can thus be followed up to the geniculate-body neurons; these neurons also degenerate (see "Transneuronal degeneration," Chap. 65).

The axons of the geniculate neurons form the geniculostriate bundle, which at first takes a forward direction around the ventricle of the temporal lobe, then swings back toward the occipital lobe, opening fanlike to end in the striate area, in the depth and lower and upper lips of the calcarine fissure. In this bundle the fibers continue to be distributed systematically, those corresponding to the macula being separated from those of the upper and lower peripheral quadrants (Fig. 443).

The Visual Area of the Cortex

The visual area in man is distributed in the depth and lips of the calcarine fissure. It has a particular cytoarchitecture with a marked line of Gennari, visible to the naked eye, from whence the name "striate area," given to this region of the cortex. It corresponds to area 17 in Brodmann's numeration and is clearly separated from the neighboring areas 18 and 19.

By correlating lesions in these areas with defects observed in the visual fields, and registering the local action potentials in the striate cortex of monkeys provoked by localized stimulation of the retina, it has been possible to establish the projection of the retina on the visual cortex. Macular fibers end on a surface that is much larger than the surface of both retinas, extending backward to the pole of the occipital lobe. The upper part of the peripheral retinas (homolateral quadrants) of both eyes ends on the superior lip down to the depth of the calcarine fissure and the lower part on the lower lip.

The visual area (17) is connected with areas 18 and 19, which cover an area three times as large as the visual area. This has been demonstrated by applying strychnine locally (Dusser de Barenne's method) and following the spread of the action potentials after stimulation of the striate area.

Functions of the area striata. The visual cortex increases in importance as species rise in the zoological scale. Thus visual acuity in fishes and birds is dependent on subcortical centers and is not abolished by removal of the cortex. In mam-

mals, vision is a cortical function; subcortical centers only mediate reflexes.

Marquis has demonstrated, by means of conditioned reflexes, that a decorticate dog can still discriminate difference in light intensity, which is dependent on the functions of the rods, but not difference in color, dependent on the cones. The centers for the rods would therefore be subcortical, and those for the cones would be in the cortex. These experiments were later confirmed in rats and cats; in monkeys the results were less clear; in man the centers for rods and cones are in the cortex, and all types of vision are lost after extirpation of the occipital lobe. In the rat, shape is not discriminated after removal of the cortex, because this process is not exclusively dependent on the rods (as brightness is) but also on the cones, which have their center in the cortex in this species.

Lesions of the Visual Pathway

Localized lesions of the visual path, provoked experimentally or due to disease, owing to systematization of the fibers along the whole visual path, produce deficiencies in limited parts of the visual fields.

Section of the optic nerve is followed by total blindness in the corresponding eye. Lesions or sections in the chiasma have different results according to their site. Thus a section through the mid-line cuts fibers from the nasal quadrants of both retinas, and vision is lost in the temporal half of the visual fields of both eyes. This condition, known as "bitemporal hemianopsia," occurs in cases of hypophyseal tumor. If the lateral fibers of the chiasma are damaged, *e.g.*, by bilateral lesions in the angle formed by the optic nerve and optic tract, the temporal half of both retinas is disconnected and the nasal half of both visual fields is lost (binasal hemianopsia).

Section of the optic tract interrupts the fibers from the temporal half of the retina on the same side and the nasal half on the opposite side. Half the visual field of each eye, corresponding to the side opposite to the lesion, is suppressed. For example, if the left optic tract is destroyed the right half of the visual fields is deficient; this is known as right lateral homonymous hemianopsia.

Fig. 444. Visual cortex.

Lesions posterior to the geniculate bodies, *e.g.*, in the occipital lobes, produce blindness in the corresponding half of both retinas but spare macular vision. This greater resistance of the macular cortex is due to the fact that it receives its blood supply from the medial and posterior cerebral arteries, while the rest of the occipital cortex is supplied only by the posterior cerebral artery (Foerster). Macular vision, however, is also spared when lesions are not produced by circulatory defects but by other means, *e.g.*, surgical extirpation or traumatic destruction. The macula therefore must be represented bilaterally, or else it has a very large area of cortical representation in the striate area which in part escapes the effect of the damaging agent. The latter hypothesis seems more probable, because macular representation extends farther than peripheral representation, not only in a posterior but also in an anterior direction.

BIBLIOGRAPHY

Adler, F. H., "Physiology of the Eye. Clinical Application," Mosby, St. Louis, 1950.

Bellows, J. C., "Cataract and Anomalies of the Lens," Mosby, St. Louis, 1944.

Bowne Eckhardt, L., J. McLean, and H. Goodell, "Experimental Studies on Headache: the Genesis of Pain from the Eye," Williams & Wilkins, Baltimore, 1943.

Christensen, G., and P. Leinfelder, A Critical Study of Lens Metabolism, *Am. J. Physiol.*, **135**, 21, 1952.

Detwiler, S., "Vertebrate Photoreceptors," Macmillan, New York, 1943.

Duke-Elder, W. S., "Text Book of Ophthalmology," Henry Kimpton, London, 1932.

Fertig, H. H., Flicker Fusion Frequency, in O. Glasser's "Medical Physics," Vol. 2, Year Book Publishers, Inc., Chicago, 1950, p. 335.

Friedenwald, J., Studies on the Physiology, Biochemistry and Cytopathology of the Cornea in Relation to Injury by Mustard Gas and Allied Toxics, *Bull. Johns Hopkins Hosp.*, **82**, 81, 1948.

Granit, R., "Sensory Mechanisms of the Retina," Oxford, New York, 1947.

Hecht, S., The Chemistry of Visual Substances, *Ann. Rev. Biochem.*, **11**, 465, 1942.

Judd, D. B., Vision: Color, in O. Glasser's "Medical Physics," Vol. 2, Year Book Publishers, Inc., Chicago, 1950, p. 1149.

Krause, A. C., "The Biochemistry of the Eye," Johns Hopkins Press, Baltimore, 1934.

Magitot, A., "Physiologie oculaire clinique," Masson et Cie, Paris, 1946.

Murray, E., Vision: Color Tests, in O. Glasser's "Medical Physics," Vol. 2, Year Book Publishers, Inc., Chicago, 1950, p. 1161.

Ogle, K. N., Optics and Visual Physiology, *Arch. Ophth.*, **45**, 684, 1951.

Peter, L. C., "The Extraocular Muscles. A Clinical Study of Normal and Abnormal Ocular Motility," Henry Kimpton, London, 1928.

Polyak, S. L., "The Retina," University of Chicago Press, Chicago, 1941.

Rochon Duvigneaud, A., "Les Yeux et la vision des vertebrés," Masson et Cie, Paris, 1943.

Scuderi, G., Permeabilità de la cornea, Editrice Minerva Medica, Turin, 1950.

Sloan, L. L., Measurement of Visual Acuity. A Critical Review, *Arch. Ophth.*, **45**, 704, 1951.

Wald, G., The Chemical Evolution of Vision, *Harvey Lect.*, **41**, 117, 1945–1946.

Walls, G. L., "The Vertebrate Eye and Its Adaptive Radiation," Cranbrook Institute of Science, Michigan, 1942.

Willmer, E. N., "Retinal Structure and Colour Vision," Cambridge, New York, 1946.

Wolff, E., "The Anatomy of the Eye and Orbit," H. K. Lewis and Co., London, 1948.

Wright, W. D., "Researches on Normal and Defective Colour Vision," Henry Kimpton, London, 1946.

CHAPTER 79

Audition

Historical note. The scientific knowledge of hearing began in the sixteenth century with the anatomical study of the ear. The simple methods used at that time had given their maximum results by the end of the eighteenth century. Most of this pioneer work was done in Italy by Vesalius (1515–1564), Ingrassia (1510–1580), Fallopius (1523–1562), Eustachius (1510–1574), and later by Valsalva (1666–1723), Morgagni (1682–1771), Cotugno (1736–1822), and Scarpa (1747–1832). To these names, those of the English anatomist Willis (1622–1675) and the Frenchman Duverney (1648–1730) should be added. In the nineteenth century the methods of comparative anatomy and histology in the hands of Corti (1822–1876), Reissner (1824–1875), and others completed the knowledge of the structure of the ear necessary for the enunciation of the first theories on the physiology of hearing.

To Helmholtz (1821–1894) goes the credit for having correlated the phenomena of audition with those studied in the physics of sound. Many important problems were thus solved and others were stated, some of which even today remain without a satisfactory answer. The best known of his works on this subject was published in 1863. He states there his theory that resonance of the basilar membrane is the mechanism by which the component frequencies of a complex sound wave are analyzed.

No important advance was made after this until the discovery in this century of electromechanical means of producing and measuring sounds of any strength and complexity, and of amplifying and recording very small electrical potentials of very short duration. The application of these techniques to the study of hearing is marked by the work of Wever and Bray (1930), who were able to reproduce faithfully sounds picked up by the ear, by amplifying and recording the electrical potentials of the cochlear nerve. In recent years knowledge has advanced at a rapid rate, as a result of the concerted efforts of physicists, psychologists, otologists, and physiologists.

The ear is a complex organ found in reptiles and higher animals. Like other sensory organs, it develops from the ectoderm of the embryo. It is extraordinarily sensitive to sound waves. It analyzes them accurately into their components and transmits them to the brain in the form of electrical potentials traveling along the axons of the cochlear nerve. These sound waves usually reach the ear through the air as very small and rapid pressure changes, caused by the vibration of the body that emitted the sound.

The ear fulfills the two essential conditions needed in an apparatus for capturing and registering sound waves: it has good resonance and damping. It can therefore faithfully reproduce a tone and prevent its persistence for a longer time than is necessary to stimulate the receptor.

This study of hearing will be divided into three parts:[1]

1. Auditory sensations.
2. The mechanics of audition.
3. Auditory pathways and nerve centers.

Disturbances caused by sound will be considered briefly; finally a general summary will be given.

AUDITORY SENSATIONS

Hearing has considerable physiologic and psychological importance. It is one of the principal means of establishing relations with the environment, particularly with other human beings. It is necessary for learning speech and

[1] The student is advised to refresh his knowledge of acoustics; he will thus find it easier to understand what follows.

song. Words are the normal means of thought and the transmission of thought from one person to another. Education and progress are therefore ultimately dependent on hearing. Auditory sensations, especially music, have a high affective tone, which is finely graded from intense pleasure to extreme displeasure. Hearing has great importance in the formation of personality and character. Deafness has a powerful influence on the psychology of the individual who suffers it, and may cause serious mental disturbance.

The Sensitiveness of the Ear. Audiometry

Sounds are vibrations of bodies which can evoke an auditory sensation in a normal ear. Tones are produced by vibrations as regular as the movement of a pendulum. Noises are produced by irregular vibrations.

Sound vibrations are transmitted by elastic bodies (solid, fluid, or gaseous) but not in a vacuum. The particles of the transmitting body vibrate longitudinally, *i.e.*, move forward and backward in the direction in which the sound wave is propagated, creating alternately areas of rise and fall in pressure, of compression and rarefaction.

A very small amount of energy is carried by an audible sound; moreover, it is only a small fraction of the energy necessary to cause the sounding body to vibrate. Thus, in electrical instruments producing sounds the proportion is 5 per cent, and in the best musical instruments only 1 per cent.

Two physical properties of sound, *frequency* and *intensity*, are of special interest to the physiologist because they are related to the perception of the *pitch* and *loudness* of a sound. Frequency is expressed by the number of double vibrations or cycles per second. Intensity is the force or strength of the sound and depends on the amplitude of the vibrations. It is difficult to express it in terms of the absolute pressure variations or energy of the sound wave. The intensity of a sound can, however, be easily expressed relatively to that of another sound. The sound taken for reference is a tone of 1,000 cycles/sec. (c.p.s.), near the threshold of the human ear (10^{-10} μw/ sq. cm.); the intensity of the sound to be compared is deduced from the voltage applied to the loud-speaker or microphone. The physical unit of intensity is called a *bel*, after Alexander Graham Bell, the inventor of the telephone. The decibel (db.), *i.e.*, one-tenth of a bel, is the unit commonly used. Formulas for number of bels and decibels are as follows:

$$N_{bel} = N_b = \log \frac{\text{Energy of sound described}}{\text{Energy of reference sound}} = \frac{E_1}{E_2}$$

$$N_{db} = 10 \log \frac{E_1}{E_2}$$

The human ear perceives a wide range of sounds; therefore, to avoid large numbers, a logarithmic scale is used, and the whole range lies within 120 db. Within certain limits the pitch and loudness of a sound can be distinguished and perceived separately.

Sensitiveness to pitch. Pitch is the quality of the sensation that permits a sound to be classified in the scale from high to low. Sounds of the same intensity (loudness) rise in pitch as the frequency is higher. The relation between pitch and frequency is not, however, absolute, because the amplitude of the sound wave (which determines its loudness) modifies the pitch. Thus if a tuning fork vibrating at a constant frequency is gradually brought nearer to the ear, the pitch of the sound becomes lower as its loudness increases.

The human ear can perceive sounds of frequencies between 16 and 20,000 double vibrations (or cycles) per second (c.p.s.), or 11 octaves. This is known as the pitch range. In music only tones between 25 and 4,800 vibrations (7 octaves) are used, and the human voice has a pitch range of 300 to 3,000 vibrations. The rat hears high tones better than the human ear and low tones not so well. Bats, when flying, emit sounds of very high pitch (supersonics), which cannot be perceived by man but which are reflected on solid surfaces, picked up by the bat's ear, and used to guide the animal when it is flying in the dark. Destruction of one ear causes serious disturbances in night flying. If both ears are destroyed, the animal cannot fly in the dark because it no longer has the means to avoid obstacles and constantly collides with them (Galambos). By means of conditioned reflexes, it has been possible to establish the pitch range of many animal species. Thus the dog has a much higher upper limit than man, and can hear the "silent" dog whistle, which emits sound waves at a higher frequency than can be perceived by the human ear.

Sensitiveness to loudness. Loudness is the quality that permits the classification of a sound as soft or loud; soft sounds are those of small intensity, loud ones those of large intensity. The relation between intensity and loudness is not, however, absolute, because the frequency (pitch) of a sound modifies its loudness. Thus if a subject

is asked to match the loudness of a sound of given frequency with that of a reference sound of 1,000 c.p.s. of given intensity, it will be seen that the intensities of the two sounds are not the same, although they are perceived by the subject as having the same loudness.

Fig. 445. Normal human auditory range. (*After Fletcher.*)

The minimum or threshold intensity of sound capable of being perceived by the human ear varies for each frequency within the pitch range. The lowest threshold (highest sensitiveness) is for tones of 2,000 to 5,000 c.p.s.; maximum audibility corresponds to tones of approximately 2,700 c.p.s. The threshold increases sharply for tones above and below these frequencies and reaches high intensities at the extreme limits of the pitch range. If the intensity is increased above the threshold, when it is sufficiently great a sensation of vibratory pressure will be perceived, accompanied by dizziness with tones of low frequency and pain with tones of high frequency. This threshold of feeling is highest for frequencies between 250 and 1,000 c.p.s., and decreases gradually for tones of lower or higher frequencies. If the auditory and sensory thresholds are plotted on a chart (Fig. 445), the former is represented by a curve convex to the abscissa and the latter by a curve concave to the abscissa; the area between both curves is the auditory range. There is a difference of 120 db. between the lowest auditory threshold and the highest tolerable intensity; the latter is a million million times the intensity of the former (Fig. 446). The auditory range is gradually reduced with age; after thirty years of age the threshold for tones of higher pitch than 1,000 c.p.s. increases.

Timbre. This is the characteristic quality of a sound which allows the ear to distinguish it from others of the same pitch and loudness. Thus the same note with the same intensity is perceived differently when played on a piano, a violin, or a trumpet. Timbre is due to the fact that most sounds are not pure single tones (as those of tuning forks) but are made up of several tones. The tone of lowest frequency and highest in-

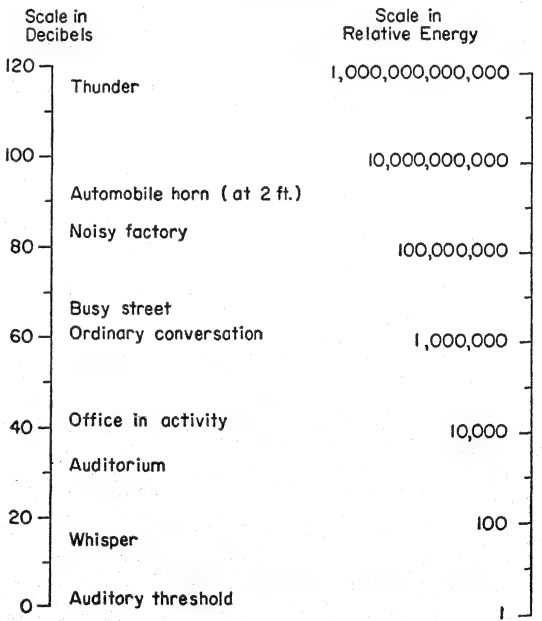



Fig. 446. Relative intensity of common noises.

tensity is known as the *fundamental tone;* others of higher frequency and lower intensity are called *overtones.* The number, frequency, and relative intensity of the overtones give the characteristic

timbre to each note. Each one of these properties of the sound can be examined by means of analyzers, and the *spectrum* of a note or sound is thus established. The spectrum of a given note varies according to its source, *e.g.*, whether it is played on a piano, a violin, or trumpet.

Difference thresholds. The ear perceives accurately small differences in intensity or pitch, but the capacity to discriminate varies for different points of the auditory range. Maximum discrimination obtains for tones within the intensity and pitch ranges of the human voice (300 to 3,000 c.p.s.); differences of 3 cycles (*e.g.*, 1,000 and 1,003 cycles) and of 1 db. are accurately perceived in this part of the auditory range. With tones of constant intensity of medium strength, 1,500 different frequencies can be perceived between the highest and lowest audible pitches. With a tone of constant frequency near the lowest threshold, 325 different intensities can be perceived. From these two figures it is not possible to deduce the total number of different tones perceived within the auditory field, because the difference threshold changes for each frequency and intensity; but by subdividing this field into small areas and determining the number of tones perceived in each of these areas, it has been estimated that approximately 340,000 different sounds can be perceived by the normal human ear.

Analytical capacity. The analysis of a compound sound consists in the discrimination of its component simple tones. This can be done by one of three methods: (*a*) by mathematical analysis (Fourier's theorem); (*b*) by means of resonators, which vibrate selectively at the frequency of each component tone; (*c*) by transforming sound by means of microphones into electrical potentials, which are then amplified and analyzed.

The human ear is capable of distinguishing the different components of a compound sound if the intensity is not too great and there are not too many overtones. This remarkable property of the ear is known as its analytical capacity.[1] Thus the ear can pick out in an orchestra not only the sound produced by each instrument, but also notes played simultaneously. The different stimuli act at the same time on the ear, which analyzes the compound result and distinguishes each component.

Audiometry. Acuteness of hearing can be tested by several methods, but in recent years the determination of the audibility curve by means of an audiometer has been generally adopted. Audiometers transform electrical energy into pure and constant tones, having a pitch and intensity that can be regulated by the operator. They are usually calibrated so that successive frequencies vary by one octave (1:2) and the intensities by 5 db. To test the normal route of sound conduction, the sound is emitted from a loud-speaker or a telephone receiver. To test bone conduction, sound vibrations are made to act on the mastoid process of the temporal bone. A tone of highest intensity (1 volt) is sounded, and then its intensity is gradually diminished down to the auditive threshold. The result is given in negative decibels below 1 volt. The operation is repeated with sounds of other frequencies.

Results are usually charted in an audiogram; frequencies are noted on the abscissae and intensities on the ordinates (Fig. 447). *Normal hearing by air conduction* is represented by a horizontal line and taken as zero. A broken line situated below the former represents *normal hearing by bone conduction.* The *auditory threshold* is represented by a curve concave to the zero line; these two lines enclose the normal *auditory field.* The different lines have been established by testing normal persons of both sexes at different ages. The results obtained by testing the subject under observation are marked on the chart; the frequencies at which there is deficient hearing and the degree of the deficiency can be seen at a glance. Loss of hearing is expressed in decibels or in per cent of normal for each tone.

This method permits the localization of auditory deficiencies in the area of the auditory field in which they occur, and it determines the pitch of the tones that must be amplified to correct them. Amplification of sounds is obtained by audiphones, instruments that are placed on the ear and consist of (*a*) a microphone that converts sound into electrical potential; (*b*) an amplifier; (*c*) a transformer that reconverts the amplified potential into a sound louder than the original one. A small battery feeds the ap-

[1] This power of analysis may be stated in the form known as Ohm's acoustical law: "Every motion of the air which corresponds to a composite mass of musical tones can be analyzed into a sum of simple pendular vibrations; to each single vibration corresponds a simple tone, which the ear can distinguish, and which has a pitch determined by the periodic time of the corresponding vibration."

paratus. Audiphones are efficient only when they are "selective," *i.e.*, when they reinforce the intensity of the tones that are poorly heard. An audiphone that increases the intensity of all audible sounds should not be used, because it increases not only the deficient tones but also

responding to the ear at which it first arrives. If the difference is greater than 1.8 msec., a double sound will be heard. The importance of the time factor can be demonstrated by listening to a tuning fork through a stethoscope with the two earpieces joined to the

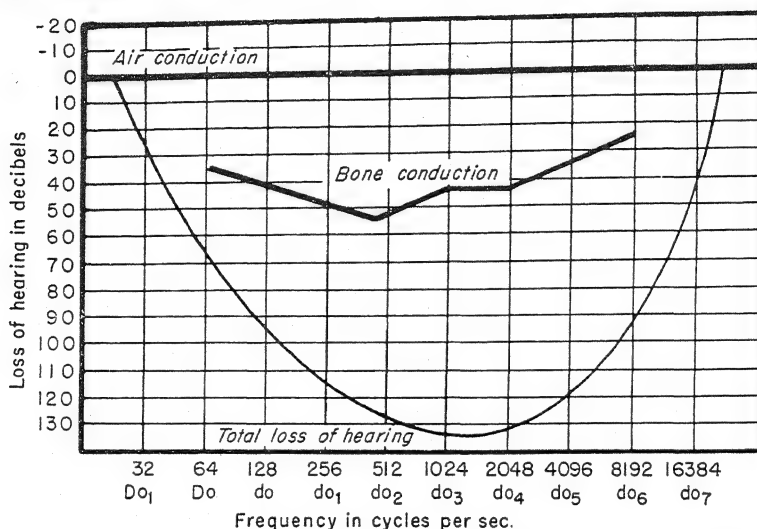



FIG. 447. Horizontal line at zero, normal hearing by air conduction. Broken line, normal hearing by bone conduction. The curve represents the auditory threshold.

all noises. Tones most needed for normal hearing are those of 1,000 to 3,000 c.p.s.; the majority of street noises are of approximately 500 c.p.s. Vacuum and crystal amplifiers are the best, because they reinforce tones of 1,000 to 7,000 c.p.s.

LOCALIZATION OF SOUND

The ear is a distance receptor, *i.e.*, it localizes sound at its source, outside the body, and can recognize the direction from which it is coming. For example, it is possible to locate the violins or a wind instrument in an orchestra and to define fairly accurately the relative position of each instrument (auditory perspective).

Sound localization is achieved by the analysis of the differences in the sound entering the two ears. Three factors concerned are of importance:

1. The difference in the time of arrival of the sound to each ear is the most important of these factors. If a sound reaches both ears simultaneously it is located as if its source were placed in the median plane of the subject. If it does not reach both ears at the same time, it will be localized on the side corresponding to the ear at which it first arrives. If the difference is greater than 1.8 msec., a double sound will be heard. The importance of the time factor can be demonstrated by listening to a tuning fork through a stethoscope with the two earpieces joined to the mouthpiece by tubes of adjustable length. When both tubes are of the same length, the sound is located in the middle plane. If one tube is shortened, so that the sound wave travels a shorter distance to arrive at one ear than at the other, the sound will appear to come from the side corresponding to the short tube (Fig. 448). The time factor is particularly important in the localization of short or intermittent sounds.
2. The different intensity of the sound as it enters the two ears is the next in importance of the three factors. The sound is referred to the side corresponding to the ear that perceives it loudest. This can be demonstrated by simultanously stimulating the ears with the same tone, but with greater intensity on one side; the sound will be referred to the side that receives the more powerful stimulus. The "listening" movements of the pinna in animals and of the head in man are such that they increase the difference in intensity of stimulation of the two ears. This factor is most effective as a means of localization with continuous sounds of 5,000 c.p.s. The sound reaches the more distant ear after it has been

partially absorbed by the "shadow" of the head; this decreases the intensity, sometimes by as much as 30 db.

3. Sound is transmitted in air by vibratory movements conveyed from one particle to another, creating a successive rise and fall in

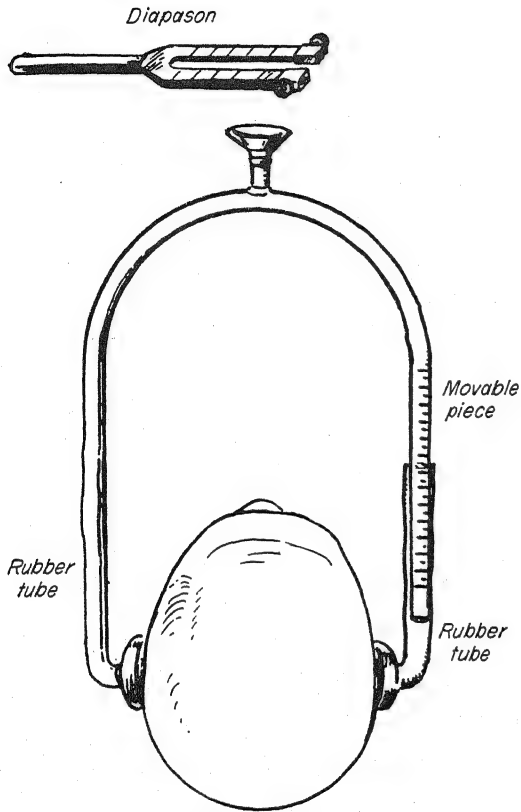



Fig. 448. Sound localization. The sound of the tuning fork will be localized in the mid-line when the arms of the stethoscope are of the same length, and on the side corresponding to the short arm when they are of unequal length.

pressure. There are therefore phases of compression and decompression. In a complete double vibration a particle is first displaced in one direction, then returns to its initial position, then is displaced in the opposite direction, and finally again returns to its initial position. These movements are similar to those of a pendulum. When the sound wave reaches both ears in the same phase, it is localized in the median plane. When it arrives at each ear in a different phase it is localized on the side at which the crest of the wave arrives first. This factor is of importance only in the localization of continuous sounds of low pitch.

Certain conditions facilitate sound localization, while others increase its difficulty. Continuous sounds are more difficult to localize than intermittent sounds, those from above than those on one or other side of the head, pure tones (especially if they are very high or very low) than noises formed by a complex of high and low tones.

These facts have been applied for the localization of airplanes and submarines. One type of apparatus has two widely separated horns, which correspond to the two ears and which are displaced so that both perceive the sound with the same intensity.

Auditory Harmonics

When the ratio of the frequency of the overtone to that of the fundamental tone can be expressed by a simple number, the overtone is called a harmonic overtone.

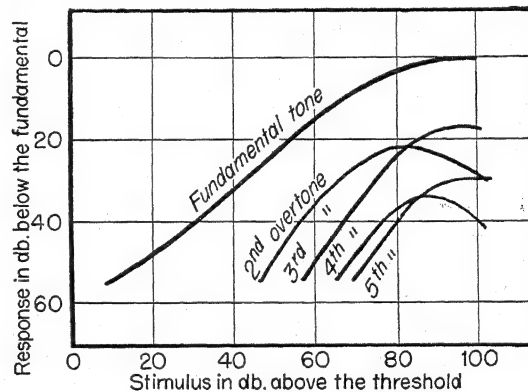



Fig 449. Microphonics of the ear of a cat stimulated by a tone of 1,000 c.p.s. Similar results are obtained with tones of other frequencies. The fundamental tone is of greater intensity than the harmonics. These are not registered when the intensity of the tone is below 50 db.; their intensity increases with the intensity of the fundamental tone. Harmonics of even number (second and fourth) decline with high intensities, but those of odd number (first, third, and fifth) remain stable. (*Obtained by Stevens and Newman, 1936, and reproduced from Stevens, S., and H. Davis, "Hearing: Its Psychology and Physiology," John Wiley & Sons, Inc., New York, 1938.*)

When a pure tone (*i.e.*, free from overtones) of an intensity greater than 45 db. reaches the ear, harmonic overtones arise in the ear. If the frequency of the fundamental is 100 cycles, the overtones will have frequencies of 200, 300, 400,

etc. The fundamental is always the most intense of the tones; the intensity of the harmonics diminishes progressively as the frequency rises (Fig. 449).

When two tones reach the ear simultaneously, a "combination tone" is heard. There are two classes of combination tones: (*a*) "difference tones"; (*b*) "summation tones." If, for example, the frequencies of the two tones are 700 and 500 cycles, there will be a summation tone with a frequency of 700 + 500 = 1,200 cycles and a difference tone of 700 − 500 = 200 cycles. The difference tone is always of greater intensity than the summation tone. Each one of the combination tones (if it is sufficiently intense) can give rise to harmonics that can also produce combination tones with other harmonics arising at the same time.

This abundance of tones arising in the ear is due to the tympanic membrane; if this membrane is destroyed, all these tones are suppressed. The tympanic membrane, owing to its asymmetry, reinforces and distorts the sound waves, producing the harmonics just mentioned when the intensity is greater than 45 db.

The existence of these overtones has been demonstrated mainly by two methods: (*a*) by the analysis of the electrical potentials of the cochlea stimulated by a fundamental of known pitch and intensity; (*b*) by sending a tone of similar pitch to the harmonic that is to be analyzed and observing the appearance of "pulsation." Telephones and phonographs reproduce only tones of high pitch, above a certain frequency, the missing harmonics being supplemented by the ear. A record engraved in this way reproduces sounds that do not differ to the ear from those of a record in which all the vibrations have been engraved.

Consonance. Beats

When two tones fall simultaneously on the ear and there is a simple ratio between their frequencies (1:2:3:4), an agreeable sensation known as "harmony" or "consonance" is perceived. Maximum consonance is obtained by two tones in unison, *i.e.*, of the same frequency (1:1). The octave (1:2), the fifth (2:3), the fourth (3:4), and other consonant intervals are employed in music. Consonance is increased by the harmonics of the fundamental tone; this occurs in a well-tuned orchestra.

Fig. 450. Interference of sound waves a and b. The resultant wave c is the algebraic sum of $a + b$.

When two tones of different but similar pitch (*e.g.*, 1,000 cycles and 1,003 cycles) reach the ear at the same time, the sound perceived undergoes periodic reinforcement and weakening; the phenomenon called "pulsation" or "beats" is produced. It is due to interference between the sound waves of the two tones (Fig. 450). When both reach the ear in the same phase they reinforce each other; when they are in opposite phases they mutually weaken their effects. Pulsation is perceived at its maximum when it occurs with a frequency of 2 to 3 per second; 6 or 7 per second are perceived as variations in pulse, and a greater frequency produces a disagreeable sensation in which the beats are no longer perceived.

Masking

A tone is masked by another simultaneous tone when the latter raises the threshold of the former (Fig. 451). An example is given by the difficulty in following a conversation in the midst of traffic noise. Masking is expressed by the number of decibels by which the threshold of one tone is increased when the other tone is simultaneously applied. The masking effect of a tone increases as its pitch is nearer the pitch of the masked tone; thus a tone of 950 cycles will have considerably more masking effect on another of 1,000 cycles than on one of 800 cycles.

It is easier to mask a tone by another falling on the same ear than by one that falls on the other ear; in this case the threshold of the masked tone must rise 40 to 60 db. There is a double

mechanism of masking: (*a*) a peripheral mechanism in the ear; (*b*) a central mechanism in the brain centers. The peripheral mechanism is the more important of the two.

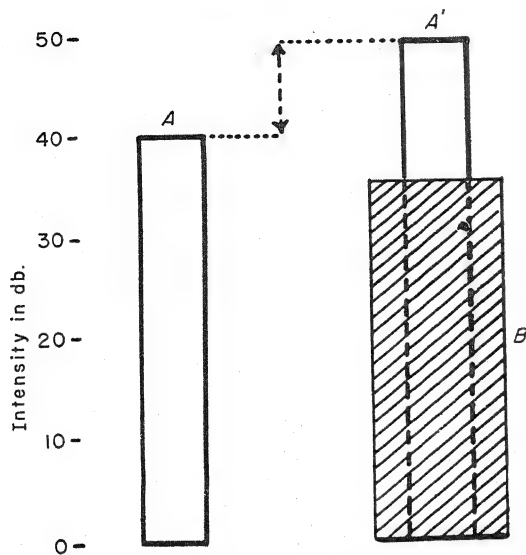



FIG. 451. Masking. *A*, auditory threshold for a single tone; *A′*, threshold for the same tone when a second tone *B* is sounded at the same time. Masking of *A* by *B* is measured in decibels by the difference between *A* and *A′*.

Fatigue

The ear is being stimulated continuously, but it is remarkable how rapidly it recovers from fatigue. There are several ways of measuring fatigue; the one most commonly used consists in determining the intensity threshold of a tone before and after the tone has been heard continuously for a certain time.

Fatigue lasts for a few seconds or minutes according to the pitch of the tone. High tones produce more prolonged fatigue than low tones. For example, if a sound of 94-db. intensity of 2-min. duration has a frequency of 100 cycles, it produces fatigue lasting 20 sec.; if the frequency is 4,000 cycles, fatigue lasts 6 min. Fatigue is not strictly specific; it is maximum for the tone used as stimulus, but it also spreads to the neighboring tones on the scale.

Rawdon Smith has observed a decrease in auditory acuteness in one ear when the other ear is stimulated. There is therefore a central factor in fatigue, dependent on the response of the brain centers, which makes the estimation of fatigue in the receptor difficult.

THE MECHANICS OF AUDITION

The ear has three parts: (*a*) the external ear; (*b*) the middle ear; (*c*) the inner ear. Not only from the structural, but also from a functional, point of view, these three parts of the ear must be considered separately but in a correlated manner.

The External Ear

The external ear consists of the lobe or pinna and the external auditory meatus. Both have a cartilaginous support, completely covered by skin. The auditory meatus is a tube 2.5 cm. long and 0.7 cm. in diameter, with a capacity of about 1 cc. Its natural frequency is 430 c.p.s. It is not straight, and the inner part is enclosed in the temporal bone. The internal end is closed by the tympanic membrane.

The pinna takes up the sound waves and contributes to the localization of sound, especially in animals with a conical and movable pinna. In man it is not so important; nevertheless, many years ago Schneider showed that hearing was impaired when the irregularities in the pinna were filled up and smoothed out with wax. The meatus amplifies the sounds and directs them toward the tympanic membrane. Reinforcement has been demonstrated by registering the intensity of the sound as it travels to the tympanic membrane. The meatus protects the eardrum, preventing by its hairs and waxy secretion (cerumen) the passage of foreign bodies (dust, insects, etc.); it also warms the air. In some aquatic animals, *e.g.*, the whale, a sphincter can close the meatus and prevent water from entering it.

The Middle Ear

ANATOMY

The middle ear, or eardrum, or tympanum, is a small cavity 1 to 2 cm. in diameter situated in the petrous portion of the temporal bone. The walls of the cavity are lined with a mucous membrane, which is ciliated except on the external wall (the tympanic membrane). The median or internal wall is a bony septum which separates the eardrum from the inner ear. In the middle of this septum there is a bony salient or promontory made by the base of the spiral lamina of the cochlea. Toward the rear of the promontory there are two openings by which the eardrum communicates with the internal ear. The upper

opening is the oval window (fenestra ovalis); the oval base or footplate of the stapes (stirrup bone) is attached to the margins of this window. The other opening—the round window—is closed by a membrane.

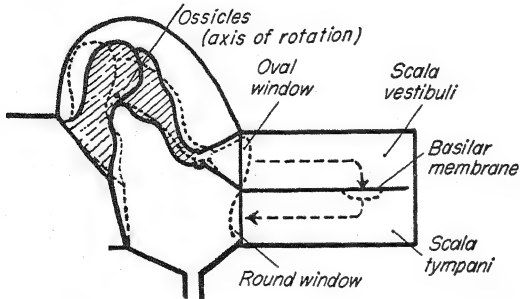



FIG. 452. Diagram of the middle and inner ear. The thick line represents the ossicle chain and the basilar membrane at rest; the broken line, the same structures when in movement. When the tympanic membrane is pushed inward by a sound wave, the base of the stapes is pushed into the oval window; at the place where the pressure wave in the perilymph crosses from the scala vestibuli to the scala tympani the basilar membrane is deformed; finally the pressure wave pushes the membrane of the round window out into the cavity of the middle ear.

The tympanic membrane is elliptical in shape; its diameters measure 9 and 10 mm. and it is 0.1 mm. in thickness. It is placed obliquely downward and inward and is slightly funnel-shaped, with the concavity toward the meatus. Seen from the auditory canal in the living subject, it has a pearl-gray color, and the manubrium of the malleus (hammer bone) can be seen through it. The manubrium is attached to the membrane like an almost vertical ray ending at its apex. The tympanic membrane is made up of a thin layer of fibrous tissue (circular and radial elastic fibers), to which it owes its properties, and is covered externally with skin and internally with the mucous membrane of the middle ear.

The eardrum is connected with the pharynx by the eustachian tube. Its posterior and upper aspects are connected with the mastoid cells.

Within the tympanum there is a chain of ossicles which extends from the tympanic membrane to the oval window. The three bones together weigh less than 55 mg., and none is more than 10 mm. in its greatest length. The manubrium of the malleus is firmly attached to the tympanic membrane, and its head is articulated with the incus (anvil bone). The tip of the long process of the incus is articulated with the head of the stapes. The base or footplate of the stapes is oval-shaped, the horizontal diameter being the longest; its surface measures 3.2 sq. mm. It is firmly attached by the annular ligament to the margin of the oval window, which it fills completely. The stapes can rotate slightly on an axis that passes through the plane separating its posterior third from the two anterior thirds; it does not have pistonlike movements (Fig. 452).

There are two small muscles in the tympanum, the tensor tympani and the stapedius. The former arises in a bony duct above the eustachian tube and ends in a tendon which turns at right angles over the cochlear process, having its insertion on the manubrium of the malleus, near its base. The latter arises in the inner wall of the eardrum and ends on the neck of the stapes. The tensor tympani is innervated by a branch of the trigeminal (fifth) nerve, and the stapedius by a branch of the facial (seventh) nerve. The fibers of both nerves arise in the same column of nerve cells situated in the pons (seventh nerve) and medulla (fifth nerve).

PHYSIOLOGY

The main function of the middle ear is to transmit to the inner ear sound vibrations which reach it through the auditory canal or the bones of the head. It can reinforce the intensity of these vibrations or protect itself and the inner ear from the harmful effects of excessively intense sounds.

The tympanic membrane. The pressure changes caused by the sound waves make the tympanic membrane vibrate, reproducing the motion of these waves and transmitting it to the manubrium of the malleus. If the membrane is gilded or a small mirror is attached to it and a ray of light is reflected on it, the motion of the membrane can be registered photographically. It will then be seen to reproduce the oscillations of a tuning fork that is made to vibrate in the proximity of the ear. It vibrates as if it were suspended by its upper margin, but owing to its shape and attachments to the temporal bone and the manubrium of the malleolus, all its parts do not vibrate uniformly. Charts have been made of the parts vibrating most at different sound frequencies. It is noteworthy that small perforations of the tympanic membrane do not alter its efficiency. The tympanic membrane is extraor-

dinarily sensitive; sound waves are perceived that cause displacements equivalent to the diameter of a hydrogen molecule. It is also aperiodic, *i.e.*, it reproduces faithfully the vibrations of sound waves of all pitches. It is well damped, and ceases to vibrate as soon as the sound waves have ended. It distorts sound waves when their intensity is more than 45 db.; this is due to its shape and the attachment of the manubrium, which prevent it from vibrating regularly in both directions. Thus harmonic overtones, previously discussed, are formed.

Functions of the ossicles. The chain of ossicles is of primary importance in the transmission of the tympanic vibrations to the oval window. If the chain is broken or is fixed by inflammatory processes, the auditory threshold rises by approximately 60 db. The movements of the ossicles are minute and not easy to study. They have been observed by means of the stroboscope or lapse-rate cinematography after removing one of the walls of the eardrum and attaching little mirrors on the ossicles or sprinkling them with powder so that they reflect light. This has been done in the human ear.[1] The chain moves as a unit; when the manubrium of the malleus is pushed inward, there is a corresponding movement of the base-plate of the stapes and vice versa. Up to 2,400 c.p.s., vibrations are transmitted simply; higher frequencies give rise to harmonic overtones.

The ossicles not only transmit the vibrations of the tympanic membrane to the inner ear, they also reinforce them, acting as a bent lever, the axis of which passes through the neck of the malleus and the body and short process of the incus. One of the arms of the lever is the manubrium of the malleus embedded in the tympanic membrane. When this moves inward, the head of the malleus and the body of the incus move outward, and the long process of the incus and the stapes move inward, *i.e.*, in the same direction as the manubrium. The area of the tympanic membrane is approximately 20 times that of the base of the stapes, and the ratio of the length of the manubrium to that of the long process of the incus is 3:2. Therefore, after discounting loss due to friction, energy applied to the tympanic membrane is increased 10 times when it reaches the oval window (Fig. 452). The base of the stapes is pushed in and out of the oval window, rotating on an axis that passes through its posterior margin, where the annular ligament is short and thick, whereas it is long and thin on the opposite side. When loud noises fall on the ear, the stapes rotates on its transverse axis and thus protects the inner ear. In summary, the ossicles constitute a lever that follows the movements of the tympanic membrane, reinforces them, and transmits them to the inner ear through the oval window.

Functions of the muscles. The two muscles in the tympanum have opposite effects on the base of the stapes. The stapedius tends to pull the stapes outward from the oval window. The tensor tympani, on the contrary, pulls the manubrium of the malleus inward; therefore it increases the tension of the tympanic membrane and pushes the stapes into the oval window. When both muscles contract simultaneously, the ossicle chain is shortened by bringing the ossicles into closer contact, and tension of the tympanic membrane increases. Contraction of these muscles is provoked by very rapid reflexes (as fast as those that provoke blinking). The usual stimuli are stimulation of the skin of, or near, the ear, and sounds of more than 45 db. Stimulation of one ear by a sound causes contraction in the muscles on the contralateral ear (Metz). Anesthesia suppresses these reflexes.

Contraction of the middle-ear muscles, owing to the greater tension of the tympanic membrane, considerably depresses the transmission of low tones, but does not modify the transmission of high tones, and may even favor it. This depression of low tones is a protective reflex, because low tones of great intensity are harmful to the inner ear.

The function of the eustachian tube. The cavity of the middle ear communicates with the exterior by means of the eustachian tube and the pharynx. Thus the tympanic membrane supports equal pressure on both surfaces and is in optimum condition for vibrating. The eustachian tube is opened by yawning, swallowing, and Valsalva's test (forced expiration with mouth and nose closed). When the pressures on both sides of the tympanic membrane are not equal, the membrane is deformed and hearing is decreased. This happens when the tube is closed by inflammatory processes (common cold), or when there is a rapid ascent or descent in an airplane, as long as the tube does not open and the pressure is not equalized. An excessive

[1] Guelke, R., and J. A. Keen, *J. Physiol.*, **116**, 175, 1952.

pressure difference may rupture the membrane, causing intense pain, hemorrhage, and in some cases severe general disturbances. If the tube is obstructed for some time (*e.g.*, owing to a cold) there is effusion into the middle ear cavity, which can be easily evacuated by puncturing the tympanic membrane.

The transmission of sound in the middle ear. Deafness. Sound vibrations are transmitted to the ear by the air in the auditory meatus (air conduction) or by the bones of the head (bone conduction). Air conduction is the usual and most important means of transmission. The impact of sound waves on the tympanic membrane makes it vibrate. This vibration is transmitted to the ossicles, and the base of the stapes moves in and out of the oval window, causing pressure changes in the perilymph of the inner ear. These changes in pressure travel up the scala vestibuli and down the scala tympani and end on the membrane of the round window, which is the only part of the walls of the inner ear that can give way to pressure (Fig. 452). The level at which the pressure wave passes from the scala vestibuli to the scala tympani varies with the wavelength of the sound wave.

when a vibrating tuning fork is placed in contact with the head. Osseous transmission was first studied by Ingrassia in the sixteenth century. Sound waves in this case usually go directly to the inner ear, but there is also transmission through the tympanic membrane owing to vibration set up in the air of the external auditory canal.

The importance of the round window is due to the elasticity of its membrane which permits equilibration of pressure within the inner ear, which is otherwise surrounded by bony inextensible walls. Davis remarks that there are two factors to be considered in the functioning of this membrane: (*a*) if its rigidity increases there is a decrease in hearing, because this rigidity depresses the vibration of the perilymph, which is fluid and therefore incompressible; (*b*) if it is protected from the direct transmission of sound waves from the middle-ear cavity, which normally damps the effect on the basilar membrane of vibrations transmitted through the oval window, auditory sensitiveness is increased. Confusion of these opposite effects on hearing has caused much disagreement as to the functions of the round window.

Up:	Tympanic membrane →	Ossicle chain →	Oval window →	Scala vestibuli
				↓
				Ductus cochlearis
				↓
Down:			Round window ←	Scala tympanica

Path of Sound Wave

The membrane of the round window is therefore submitted to pressure on both its sides whenever a sound wave falls on the ear—the pressure transmitted through the ossicles and the perilymph of the inner ear on one side, and the pressure transmitted directly by the air in the cavity of the middle ear on the other. These pressures act on the membrane in opposite directions, and the former is by far the greatest. When the tympanic membrane and the ossicle chain are destroyed, the sound waves act directly on the round window and the pressure changes in the perilymph travel up the scala tympani and down the scala vestibuli, *i.e.*, in a direction opposite to the normal.

There is a third possibility; sound waves are transmitted by the bones of the cranium, *e.g.*,

Pathological (inflammatory reactions, effusion, pus, etc.) or experimental disturbances of the external or middle ear cause different degrees of deafness owing to detects of *transmission*. The functional loss of the ossicles diminishes hearing by 45 to 60 db. In this type of deafness perception of low tones is usually more impaired than that of high tones, and considerable improvement in hearing can be obtained by means of hearing aids. Lesions in the inner ear cause deafness due to deficiencies in *perception*. This type occurs in old people, who have difficulty in perceiving high tones, and in subjects exposed to loud noises during prolonged periods (boilershop workers, aviators, etc.). Hearing aids are uselsss in these cases. A third type of deafness is due to lesions in the auditory nerve paths and

centers. Several tests (Weber, Rinne, Schwabach) have been devised to differentiate these types. These tests are based on the changes undergone by air and bone transmission of sound vibration. For example, in transmission deafness, only the air route functions deficiently; therefore a tuning fork is not heard when placed in the air near the ear, but its vibrations are transmitted normally if it is placed on the mastoid process of the temporal bone. In perception deafness the tuning fork is not heard, whether it is placed in the air or is in direct contact with the head.

Protective action. The middle ear has a protective action, defending the inner ear from the effects of excessively intense sounds. This is obtained by two mechanisms:

1. The air within the cavity of the middle ear acts as an elastic cushion and diminishes the vibrations of the tympanic membrane and the ossicle chain. This effect disappears if the tympanic membrane is perforated, and it is offset by the communication with the exterior through the eustachian tube, which tends to equalize the pressure on both sides of the tympanic membrane.
2. Vibration of the ossicle chain is regulated by reflex contraction of the auditory muscles.

Fig. 453. Section through the cochlea, showing the organ of Corti.

The Inner Ear

The inner ear and the auditory pathway are the essential parts of the auditory apparatus. Destruction of the cochlea causes irreparable deafness on the side on which it occurs. If the cochlea alone is destroyed, there are no disturbances in the maintenance of body equilibrium; on the other hand, if the labyrinth is destroyed without damage to the cochlea, hearing is not disturbed.

ANATOMY

The labyrinth is made up of a series of cavities in the petrous portion of the temporal bone. These cavities (the osseous labyrinth) contain a series of membranes (the membranous labyrinth) and are filled with a fluid, the perilymph. There are three cavities: (*a*) the vestibule, the center cavity, which measures 5 mm. in diameter and is connected to the middle ear by the oval window; (*b*) the three semicircular canals, placed behind and above the vestibule, which have nothing to do with hearing; (*c*) the cochlea, in front of the vestibule, which is the auditory receptor.

The cochlea is a tube coiled in a spiral fashion around a bony axis, conical in shape, known as the modiolus or columella. The whole resembles the shell of a snail, whence its name. It is

approximately 5 mm. high and 9 mm. at its base. The tube has two and a half coils in man, three in the dog and cat, and four and a half in the guinea pig. Inside the tube there is a bony shelf inserted on the modiolus which does not reach the outer wall of the tube; this is the spiral lamina (Fig. 453). The free border of this lamina is joined to the wall of the tube by a stout connective-tissue membrane (the basilar membrane). The tube is thus divided into two passageways. The upper half is again divided in two by Reissner's vestibular membrane. This membrane stretches from the spiral lamina, near the insertion of the basilar membrane, to the outer wall of the tube, some distance above the insertion of the latter. The cochlear tube is thus divided into three ducts: (*a*) the upper, known as the scala vestibuli, which extends from the vestibule to the apex of the cochlea, where it communicates by means of a small orifice, the helicotrema, with the next; (*b*) the lower duct, known as the scala tympani, which ends in the round window; (*c*) between these ducts, the cochlear duct, or scala media, closed at both ends, but communicating by a minute canal (Hensen's duct) with the sacculus. The cochlear duct is separated from the scala vestibuli by the membrane of Reissner, and from the scala tympani by the basilar membrane which arises between the oval and the round window; the diameter of its first spiral turn is 6 mm., and that of the second is 4 mm. (Fig. 454). These ducts are filled with fluid, which is called endolymph in the cochlear duct and perilymph in the scalae tympani and vestibuli.

FIG. 454. The internal ear, showing the osseous and membranous labyrinth.

The mechanism for selectively capturing the sound waves and converting them into nerve impulses is found in the scala media or cochlear duct. The length of the basilar membrane is 30 mm.; its width increases gradually from 80 μ at the base to 500 μ at the apex of the cochlea. This difference in width is a sign of the localiza-

FIG. 455. Organ of Corti

tion in different parts of the membrane of the receptive mechanisms for tones of different frequency. There are fibers of many different lengths in the basilar membrane. Helmholtz was the first to consider them the receptors of sound waves.

On the inner part of the vestibular aspect of the basilar membrane there is a complex structure known as the spiral organ of Corti (Fig. 455). It is formed by highly developed epithelial cells and fibers of the acoustic nerve. There are two pillars formed by long rodlike cells (the rods of Corti), an internal and an external pillar. Separated at their base on the basilar membrane, they come together at their apex, covering a triangular tunnel (the inner tunnel of Corti). The lumen of this duct has a diameter of 50 μ at the base of the cochlea; it widens gradually up to 100 μ at the apex. On the internal aspect of the internal pillar, and on the external aspect of the external pillar, there are the inner and outer hair cells, each one of which has up to 100 hairs of different lengths, forming a tuft which touches the tectorial membrane. These cells are not placed directly on the basilar membrane, but are supported by Deiters' phalangeal cells. There are approximately 3,500 internal hair cells distributed in a single row; and some 20,000 external hair cells in three rows. The internal hair cells are 8 μ in length, and each one receives only one nerve fiber; the external hair cells are 12 μ in length and each one receives several nerve fibers.

The inner part of the organ of Corti is covered by the tectorial membrane. This structure is attached to the spiral lamina at one end and is free at the other. It is formed by a homogeneous substance in which there are fine colorless fibers. Its inferior aspect is in contact with the hairs of the hair cells. The significance of the tectorial membrane is still unknown.

PHYSIOLOGY

Sound vibrations, transmitted through the middle ear, are converted in the cochlea into nerve impulses which travel along the acoustic nerve. Pitch, loudness, and timbre have their origin in the cochlea. The intimate mechanism of cochlear function is unknown, but the discovery of cochlear microphonics and the application of new techniques have given valuable information.

The microphonic response of the cochlea. In 1930, Wever and Bray[1] reported that a telephone or loud-speaker connected by means of electrodes to the acoustic nerve of a decerebrate cat reproduced sounds and even words whispered into the animal's ear. This fact was later confirmed in several other species. At first this effect was attributed to action currents in the acoustic nerve, but later Adrian suggested, and Saul and Davis demonstrated, that it had its origin in the cochlea.

Definition. The microphonic response of the cochlea (or simply cochlear microphonics), so called because of its resemblance to the electrical phenomena provoked by sound waves in microphones, consists in electrical potentials caused by the stimulation of the cochlea by sound. These potentials reproduce the frequency and form of the sound. Their exact nature and significance are still unknown. They precede the action potentials in the acoustic nerve. The sequence, therefore, is as follows: sound waves → microphonics → acoustic-nerve potentials.

Properties. Cochlear microphonics can be amplified and registered by a cathode-ray oscillograph, or they can be reconverted into sound. They can be detected by electrodes placed on different parts of the head, because they are transmitted to the tissues surrounding the cochlea, *e.g.*, the meninges, the acoustic nerves (especially near the cochlea), the cochlea itself, and the round window. The polarity of the current varies according to the placement of the electrodes and the phase of the sound wave. Variations in polarity in the guinea pig are given in Table 110.

Table 110. Variation in Polarity of Microphonics in Relation to Pressure Changes in the Perilymph

Cochlea		*Polarity of microphonics*	
Movement of stapes	*Pressure change*	*Oval window*	*Round window*
Enters	+	+	−
Withdraws	−	−	+

Source: Stevens, S., and H. Davis, "Hearing: Its Psychology and Physiology," Wiley, New York, 1938.

A remarkable characteristic of cochlear microphonics is the faithful reproduction of the

[1] Wever, E., and C. Bray, *Proc. Nat. Acad. Sc.*, **16**, 344, 1930.

original sound, except that loud sounds have added onto them the auditory harmonics arising in the middle ear. Thus an observer placed in a distant room can hear (by means of a telephone connected with an apparatus that captures and amplifies the cochlear microphonics) what is said into the ear of an animal, and can even recognize the voice of the person speaking.

Microphonics show the same sensitiveness to frequency and intensity as auditory phenomena. Sometimes differences are found because of deficiencies in the recording apparatus or the placing of the electrodes; these differences disappear when an adequate instrument is used and the electrodes are suitably placed. The voltage of microphonic potentials increases with the intensity of the sounds, but very loud ones provoke a decrease of as much as 30 per cent of the maximum voltage. This decrease persists for several days; transitory or permanent lesions of the organ of Corti have been observed at the same time.

There is a very short latent period (0.1 msec.) between the arrival of the sound and the microphonic response. Between the microphonic response and the beginning of the acoustic-nerve action potential there is a much longer interval (0.7 msec.). At the beginning and end there are three or four ample vibrations, different from the others, with a periodicity of 700 to 2,000, which is the same as the periodicity of the ear. These vibrations are due to inertia and incomplete damping of the ear.

Severe oxygen deprivation causes cochlear potentials to decrease, but there is no decline when the gas mixtures contain more than 4 per cent oxygen, equivalent to altitudes up to 40,000 ft.[1] The reduction in the amplitude of the potential seems to be a direct effect of hypoxia on the organ of Corti. Repeated periods of hypoxia are followed by complete restitution of the cochlear potentials, but the action potential of the acoustic nerve gradually disappears. This method has been used to eliminate the action potential without affecting the cochlear potential.

The microphonic effect is not suppressed by anesthesia, and it persists for several hours after death, when it diminishes gradually. It has, however, been observed even 5 hr. after the animal had died.

Site of origin. Two facts seem to have been well established: (*a*) cochlear microphonics arise in the organ of Corti; (*b*) microphonics due to high tones arise near the base of the cochlea, and those due to low tones arise near the apex. It is still a subject of discussion whether the hair cells are or are not the precise locus of origin of cochlear microphonics.[1]

Microphonics are not observed in animals with congenital deafness and lesions in the organ of Corti. This has been observed in "dancing" guinea pigs with disturbances in the labyrinth and in albino cats and dogs; it is a good proof that microphonics arise in the organ of Corti.

High and low tones stimulate the production of microphonics in the basal parts of the cochlea and in the parts near the apex, respectively. This has been demonstrated in several ways. For example, localized lesions can be produced in the organ of Corti by a sustained loud pure tone, or by injecting cocaine or concentrated NaCl into the round window, or by extirpating the apical coils of the cochlea. In each case there is deafness, and microphonics are decreased or suppressed, for a limited number of tones, the frequency of which varies with the locus of the lesion.

These experiments destroy all the elements of the organ of Corti and do not allow any conclusions to be drawn as to what part of this organ is responsible for the production of microphonics. In other experiments, fibers of the acoustic nerve have been cut near the internal meatus. This operation causes retrograde degeneration of the spiral ganglion and of the fibers that innervate the hair cells. Microphonics persist as long as the hair cells themselves do not degenerate. Davis suggests that the hair cells are comparable to quartz crystals in the production of piezoelectric effects; *i.e.*, they respond to mechanical deformation by the production of an electric current. According to this hypothesis, movements of the basilar membrane would exert pressure or tension on the hair cells and thus give rise to electrical potentials.

Significance of microphonics. Microphonic potentials are a valuable sign that sound has stimulated the organ of Corti and of the normal or abnormal response of this organ. They are therefore increasingly used in the experimental study of hearing, since they reproduce the sound

[1] WEVER, E. G., *et al.*, *Am. J. Physiol.*, **159**, 199, 1949.

[1] The lateral lines of fishes have hair cells which have a microphonic response to water currents (JELOF, R., A. SPOOR, and H. DE VRIES, *J. Physiol.*, **116**, 137, 1952).

waves with the same intensity threshold and tonal field, and moreover serve to localize the part of the cochlea at which a lesion is situated. Microphonics, nevertheless, do not give any information on the nature of auditory sensation, which depends on the central nervous system; yet section of the acoustic nerve, which produces complete deafness on the corresponding side, leaves the microphonics unaltered until the organ of Corti, especially the hair cells, degenerates.

The true significance of microphonics is still unknown. Some workers believe that they are the normal stimulus of the acoustic nerve endings, but there is no definite proof that this is so.

The human cochleogram. Microelectrodes have been placed in the round window of human subjects and microphonics recorded when the ear was stimulated by different sounds. They are similar to microphonics recorded in animals.[1]

Summating potential. Another electric potential, interpolated between the microphonics and the acoustic-nerve action potential, has been described. It appears after two or more sound vibrations have reached the cochlea; hence its name. This potential apparently arises in the hair cells, at the pole where the acoustic-nerve fibers end It seems to represent the local excitatory process that initiates the auditory-nerve impulse.[2]

Tonal localization in the basilar membrane. The problem of what part of the auditory system serves to distinguish the pitch of a tone has been the subject of much discussion for many years. Some observers have maintained that it is a function of the cochlea, others that it is a function of the central nervous system. Rutherford believed that the whole basilar membrane vibrated to sound, like the membrane of a telephone, and transmitted impulses to the nerve centers, where they were analyzed. This "telephone theory" can no longer be accepted because, among other reasons, there is no close relationship between the frequency of a sound wave and the action potentials in the acoustic nerve.

Helmholtz's "cochlear" or "resonance" theory, proposed over 85 years ago, is more in agreement with modern facts, although it must be modified to a certain extent. According to Helmholtz the cochlea, and not the nerve centers, analyzes the pitch of sound. The basilar membrane has fibers of different length which vibrate selectively to tones of different frequency, in the same way that the strings of a piano vibrate in response to a sound of the same periodicity. The nerve centers would recognize pitch by localizing the place of the basilar membrane that is stimulated.

It is now generally admitted that tones of each frequency stimulate selectively a definite place in the basilar membrane, which by vibrating stimulates the corresponding nerve endings of the acoustic nerve, and nerve impulses are transmitted to the auditory nerve centers. There is as yet no agreement as to what element (fibers or other structures) in the membrane is responsible, and still less is known about the mechanism by which the sound wave acts.

Proofs of localization. There are many proofs of localization of pitch in the different parts of the basilar membrane. Anatomical, experimental, and pathological observations give concurrent results. Thus hardness of hearing in old age is especially notable for high tones; this coincides with degenerative lesions in the basal coil of the organ of Corti, and has been reproduced experimentally by injecting cocaine or sodium chloride into the round window, thus destroying the basal coil and producing deafness to high tones. Guinea pigs that are stimulated with a pure tone of high intensity for a long time become deaf to that tone, as is demonstrated by the conditioned reflex technique and the absence of a microphonic response to the tone. This effect coincides with degenerative lesions in a definite place of the basilar membrane, which varies according to the pitch of the tone.

These methods do not allow precise localization, because the lesions extend to a relatively large part of the organ of Corti, and there is deafness to a large number of tones. The advent of microphonics has been an important aid to these studies. Culler[1] trephined the osseous cochlea in guinea pigs in 25 different places between the base and apex of the cochlea without causing damage to the membranous structure. He was therefore able to record the threshold of microphonics for tones of different pitch at each one of the 25 places. He observed that there was an optimum place for each pitch,

[1] Lempert, J., *et al.*, *Arch. Otolaryng.*, **51**, 307, 1950.
[2] Davis, H., *et al.*, *Proc. Nat. Acad. Sc.*, **36**, 580, 1950.

[1] Culler, E. A., *Ann. Otol., Rhin. & Laryng.*, **44**, 809, 1935.

and that high tones were placed near the base of the cochlea and low ones near the apex. Walzl and Bordley[1] were able to produce localized lesions in the basilar membrane of the cat, without damage to the delicate membrane of Reissner. They observed an increase in the threshold for tones of definite pitch, which varied according to the place where the lesion was made.

Results. The facts so far described are in agreement with the theory that each place on the basilar membrane is particularly susceptible to tones of a certain pitch. Maps have been constructed which show that tones of low pitch act near the apex of the cochlea, those of high pitch near the base, and intermediate ones (*e.g.*, those of 2,700 cycles) in the middle (Fig. 456). The areas for the different low tones are placed near each other; those for the high tones are more widely spaced. It is necessary to use tones of threshold intensity for the precise localization of the area for a certain pitch; stimulation by louder tones spreads to neighboring areas.

Response of basilar membrane to intensity of sound. The basilar membrane is an elastic structure which vibrates on the reception of sound waves. These vibrations have been observed directly in animals and in the human cadaver by stroboscopic observations with the microscope. The site of vibration varies with the frequency of the sound wave, and the area vibrating becomes larger as the intensity of the sounds increases. The voltage of the microphonics is related to the size of the area of basilar membrane which is being stimulated; it increases proportionally to the logarithm of the intensity of the tone used as stimulus (Weber and Fechner's law). With intensities above 50 db., harmonic overtones appear. These overtones must arise in the cochlea, not in the middle ear, because destruction of the ossicles or the whole tympanum does not suppress them.

FIG. 456. Map of human and guinea-pig cochleas, showing the localization of pitch reception on the basilar membrane. Figures represent c.p.s. of sounds received. (*Stevens, S., and H. Davis, "Hearing: Its Psychology and Physiology," Wiley, New York, 1938.*)

Electrical stimulation of the cochlea. Electrophonic effect. If one electrode is placed on an ear and another on an arm, the passage of alternating currents of 200 to 10,000 c.p.s. stimulates the ear. Sounds are heard, with a pitch directly related to the frequency of the current. Stevens showed that if a radio set is directly connected with the head, it is possible to hear the music and words, although with some distortion. If the current is too intense, *i.e.*, the equivalent of more than 20 db., the subject may present symptoms of shock.

This electrophonic effect has been interpreted as an inverted microphonic effect. In this case electrical potentials are perceived as sound, while in the case of microphonics sound gives rise to electrical potentials. The tympanic membrane modifies the electrophonic effect; the normal ear perceives the tones as one octave higher, while subjects who have lost this membrane perceive the tone with its true pitch.

The electrophonic effect has its origin in the cochlea. It does not exist in subjects in whom the cochlea is damaged and in whom the acoustic nerve is stimulated directly; they hear only noises. It is possible that these facts may be usefully applied in the treatment of middle-ear deafness.

The dynamics of the inner ear. The facts so far discussed can be used to form an idea of how the inner ear functions when a sound falls on the ear. The vibrations of the plate of the stapes provoke pressure changes in the perilymph of the inner ear. The osseous labyrinth is inextensible; therefore the pressure waves travel up the scala vestibuli and down the scala

[1] WALZL, E. M., and J. BORDLEY, *Am. J. Physiol.*, 135, 351, 1942.

tympani, and end on the membrane covering the round window. The point at which the pressure wave passes from one scala to the other varies with the pitch of the tone. If the pitch is low, the wave travels up to near the apex of the cochlea; if it is high, it passes to the scala tympani near the base. On crossing from one scala to the other, the waves cross the scala media and cause the basilar membrane and the organ of Corti to vibrate at that place and to stimulate the corresponding nerve endings of the acoustic nerve. The organ of Corti gives rise to the microphonics, perhaps owing to deformation of the hair cells (Figs. 452 and 453). A loud tone is distinguished from a soft tone because it causes vibration in a wider area of the basilar membrane.

AUDITORY PATHS AND NERVE CENTERS

Anatomy

From the cochlea to the temporal cortex, auditory impulses pass through four neurons (Fig. 457). The neurons of the first order are situated in the modiolus and form the spiral ganglion of the cochlea. They are bipolar cells, 25,000 to 29,000 in number. Each has a short peripheral process ending on the internal or external hair cells, and a long central process which forms part of the cochlear division of the eighth nerve. They pass out of the temporal bone through the internal meatus and end in the cochlear nuclei. These are two groups of cells (second-order neurons) situated in the restiform body, where the latter turns toward the cerebellum. The dorsal cochlear nucleus forms the tuberculum acusticum on the dorsal aspect of the restiform body; the ventral cochlear nucleus, or ganglion ventrale, is situated in the ventrolateral aspect of the restiform body. The ventral cochlear nucleus emits fibers which cross the mid-line, forming the trapezoid body, and then turn rostrally to form the bundle known as the lateral lemniscus, which ends in the inferior colliculus. The dorsal cochlear nucleus sends out fibers toward the mid-line, running under the floor of the fourth ventricle (striae acusticae). After these fibers decussate, they sink into the reticular formation and join the lateral lemniscus, also ending in the inferior colliculus. Part of the secondary auditory path does not decussate but ascends in the homolateral lemniscus.

Third-order neurons of the auditory path are found in the superior olivary nuclei, the nuclei of the trapezoid body, the nucleus of the lateral lemniscus, and the inferior colliculus. Their efferent fibers form the brachium of the inferior colliculus and end in the medial geniculate body. The fourth-order neurons in this body send fibers to the temporal cortex through the internal capsule. The medial geniculate body receives fibers from other sensory pathways and from the cortex. Probably it acts as a coordinating center. It is also systematized, and each point of the organ of Corti has its corresponding point in the geniculate body, but this systematization is not so precise as that of the lateral geniculate body in the visual path.

Fig. 457. Auditory path in the central nervous system. (*After Rasmussen.*)

The auditory cortex in the temporal lobe is divided into two parts:

1. Poljak's "nuclear area" (areas 41 and 42 of Brodmann) in the superior temporal gyrus, deep in the lower lip of the sylvian fissure. It measures only 8 × 4 mm., but it receives nearly all the auditory fibers. It is formed by large pyramidal cells, and

numerous fibers of which a deep layer runs parallel to the cortical surface.
2. The remainder of the temporal cortex (area 22 of Brodmann) is formed by small pyramidal cells and is not limited by a deep fibrous layer. It receives fibers from the auditory pathway and nonauditory fibers.

Physiology

Electrical potentials in the auditory nerve. When the ear is stimulated by sound, action potentials arise in the acoustic nerve, corresponding to nerve impulses sent out from the cochlea. They should be registered at some distance from the inner ear; otherwise they are covered by the cochlear microphonics. The mechanism by which they arise is not yet fully understood, but several facts of interest in this respect are fairly well known.

Variations in pressure in the perilymph and endolymph can no longer be considered as stimulating the nerve endings directly; the mechanical theory of stimulation must be discarded. Pressure changes in the endolymph caused by sound vibrations distort some element in the organ of Corti. All the evidence points to the hair cells as the receptors and indicates that microphonics represent the electrical response of these cells. A local excitatory state then arises, probably at the foot of the hair cells, where they are in contact with the auditory-nerve endings. The summating potential is the electrical manifestation of this phenomenon. After a relatively long latent period (0.7 msec.), the propagated nerve impulse is fired off.

A chemical mediator may play a part in this process. There is no evidence that acetylcholine is released when the inner ear is stimulated by sound waves. Cholinesterase, however, has been found in the endolymph and perilymph in sufficient amounts to play a physiologic role. Moreover, cholinesterase inhibitors, *e.g.*, physostigmine, prolong the latent period of the acoustic potentials.[1] More facts are needed before it can be concluded that a chemical mediator, probably acetylcholine, is a factor in the excitation of the receptor or the transmission of the excitatory state from the receptor to the nerve endings.

The action potentials of the cochlear nerve are similar to those of other sensory nerves, but they are very sensitive to anoxia and cease soon after the animal dies. They begin 0.8 msec. after the arrival of the sound wave. The first oscillation is always negative, while the first microphonic potential may be negative or positive. It lasts 1 msec., as in other myelinized fibers of similar diameter, and the refractory period is also of 1 msec. The speed of conduction is 30 m. per sec., as in the optic nerve and in tactile fibers. The threshold varies with the distance of the active fiber from the electrode.

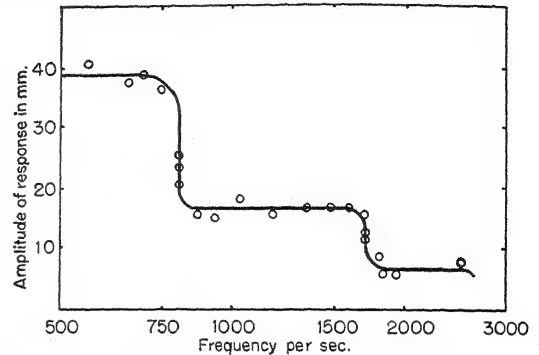



Fig. 458. Amplitude of the action potential of the auditory nerve of the cat related to the pitch of the stimulating tone. The first drop occurs around 900 cycles, when half the fibers are stimulated alternately; the second drop occurs around 1,800 cycles, when one-third of the fibers are stimulated in rotation. (*Stevens, S., and H. Davis, "Hearing: Its Psychology and Physiology," Wiley, New York, 1938.*)

Galambos and Davis[1] have registered the action potentials of single fibers. Each fiber is stimulated by tones of a limited number of frequencies (pitch) when the sound is near the threshold intensity. As the intensity increases, the fiber responds to a larger number of frequencies.

Relation of action potential to sound. Action potentials in the auditory (cochlear) nerve are clearly related to the pitch of the stimulating tone. Stimulating with tones of increasing frequency, it has been observed that, up to 3,500 to 4,000 cycles, for each vibration there is a corresponding nerve impulse. Above this frequency the individual impulses can no longer be recognized. The maximum amplitude for each tone drops to one-half for tones of 900 to 1,800 cycles, to one-third for tones up to 2,700, to one-quarter for higher tones, after which there is no further decrease (Fig. 458). The

[1] Gisselsson, L., *Acta oto-laryng.*, Suppl. 82, 1950.

[1] Galambos, R., and H. Davis, *J. Neurophysiol.*, **6**, 39, 1943.

explanation of this fact is given by alternation and rotation in the activity of the different fibers (Fig. 459).

Up to 900 cycles, all the fibers of the nerve that transmit at this frequency conduct impulses at the same frequency as the sound waves, and the amplitude of the action potential is the sum of the axon potentials of all these fibers. The refractory period of the fibers does not permit them to respond to frequencies greater than 900 cycles. Tones of higher frequencies stimulate alternately one-half of the fibers, with the following results: (*a*) each fiber responds to half the stimuli, *e.g.*, if the frequency of the tone is 900 cycles, some fibers will respond to half these stimuli (450) and others to the other half; (*b*) at a given moment only half the fibers are active; (*c*) the action potential therefore is one-half of the maximum potential for lower frequencies. When the frequency increases to 1,800 cycles, owing to the refractory period, only one-third of the fibers responds, and excitation rotates through three sets of fibers; therefore the voltage of the action potential is reduced to one-third. With frequencies of 2,700, excitation rotates through four sets of fibers, and the voltage of the action potential is reduced to one-quarter. In summary, as the frequency (pitch) of the tones rises, owing to the refractory period of the fibers of the cochlear nerve, each sound wave provokes simultaneous nerve impulses in a smaller number of fibers, and the number of sets of fibers activated in rotation increases.

Fig. 459. Diagram of action potentials in the fibers of the auditory nerve in response to stimulation with tones of different pitch. Alternation and rotation.

The fibers of the auditory nerve are systematized; tones are transmitted by different fibers according to their pitch. Therefore by placing the electrodes on different parts of the nerve, the threshold for tones of different pitches can be varied.

If the ear is stimulated with a tone of constant pitch, as the intensity (loudness) increases a larger number of fibers are brought into activity and the action potential increases. This effect is due to the larger area of the basilar membrane which vibrates as the intensity of the sound increases.

"Masking," *i.e.*, decrease in loudness of a tone when it is perceived simultaneously with another tone, can be demonstrated by recording the action potentials of the auditory nerve. The action potential provoked by a tone diminishes when the ear is stimulated by a second tone. This effect is especially marked if the intensity of the first tone is close to the threshold. The second tone sets up impulses in the same fibers and leaves a refractory period, so the "line is busy" during certain intervals and cannot be used by the first tone. This phenomenon is not

observed in the microphonic potentials, which summate.

Hartridge and his collaborators have observed that if, while a tone is being perceived, the phase of the tone is changed by half a cycle, it appears to be interrupted for a brief interval, and is then

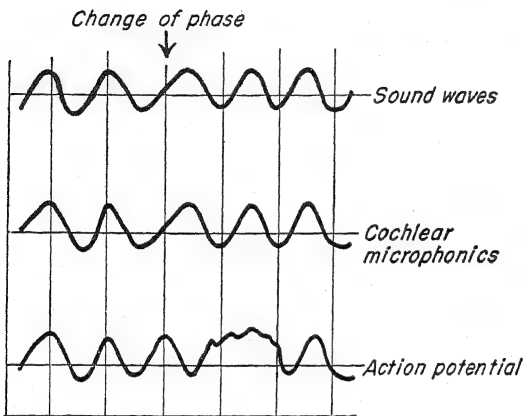



FIG. 460. Diagram of simultaneous records of sound waves, cochlear microphonics, and auditory-nerve action potentials. At the arrow the phase of tone is changed by half a cycle.

again heard. The same phenomenon is observed in a resonator, a fact that has been brought forward in support of Helmholtz's theory of hearing. By registering simultaneously the vibrations of the tone, the cochlear microphonics, and the action potential of the auditory nerve, it is seen that the microphonics reproduce faithfully the change of phase, but the action potentials are interrupted and then again respond rhythmically (Fig. 460). According to Davis, these observations show that the ear is an aperiodic system, and the delay is caused by the nerve; therefore the cochlea does not behave as a resonator.

Differences between action potentials and cochlear microphonics. Action potentials and cochlear microphonics are two separate phenomena, both due to the arrival of sound waves at the cochlea. Confusion between them has arisen owing to the proximity of the structures where they arise, the remarkable similarity of their thresholds, and the fact that microphonics are sufficiently intense to "cover" the action potentials if due precautions are not taken. At one time they were thought to be only one phenomenon, but later their separate individualities and significances were established. The differences between them are summarized in Table 111.

Auditory paths and centers. Action potentials in the auditory tract (Fig. 457) have been recorded by means of special electrodes. This method has shown that the impulses travel to the cortex on the same, and on the opposite, side of the ear in which they arise. The delay increases considerably at each synapse; it has therefore been used to establish the number and situation of the neurons in the path. Between the cochlea and the auditory cortex there are four neurons. In neurons of the first order, *i.e.*, those of the auditory nerve, there is perfect synchronization of the impulses and sound waves up to a frequency of 3,500 c.p.s.; in the second-order neurons, synchronization is achieved only up to 2,500 cycles; in the third-order and fourth-order neurons it exists only for tones of low frequencies (less than 500 cycles). The higher neurons are also more susceptible to anesthesia than the lower ones.

Table 111. Differences between the Action Potentials of the Auditory Nerve and Cochlear Microphonics

	Cochlear microphonics	*Action potentials*
In relation to sound		
Latent period, msec	0.1	0.8
Form of sound wave and electrical potential wave	Same	Different
Synchronization with tone	Perfect	Only up to 3,000 cycles
Masking	No	Yes
Change of phase in sound	Followed	Not followed (silent interval)
Susceptibility to anesthesia, cold, ischemia, death	Little	Much
Polarity	Variable	Negative
Place of registration	Several	Auditory nerve
Refractory period	No	Yes

There is a certain degree of systematization in the auditory nuclei related to tonal frequency. The ventral cochlear nucleus receives fibers from the base of the cochlea. The medial geniculate body is also laminated. Ades, Mettler, and Culler[1] have shown in the cat that destruction of one layer in the medial geniculate body causes deafness to tones of certain pitch. There is also systematized projection of the medial geniculate body on the homolateral auditory cortex. Localized destruction of the medial geniculate body causes localized degeneration in the cortex.

The auditory area in the temporal cortex comprises two parts with recognizable differences in structure. The following methods have been used in the study of auditory cortical functions in cats, dogs, monkeys, anthropoids, and man.

1. *Anatomoclinical observations.* Observations in patients with disturbances in hearing are correlated with lesions found at autopsy or in the course of surgical operations. Thus, in cases of tumors in the temporal lobe there is a decrease in auditory acuteness, localization of sound is deficient, and nonexistent sounds are heard (auditory hallucinations). The disturbances are present in both ears, but more marked on one side. Destructive lesions of the temporal cortex cause aphasia (see Chap. 80).
2. *Extirpation of the temporal cortex* causes disturbances which have been studied in animals by conditioned reflexes to certain tones. If the cortex is removed on one side only, there is good compensation, but the threshold rises by 2 to 5 db. If the cortex is removed on both sides, the auditory threshold rises by 75 db., but some degree of hearing remains. The bilateral distribution of the auditory pathway has been demonstrated by removing the temporal cortex on one side and destroying the cochlea on the other; in this case the threshold rises by 15 db.
3. *Electrical stimulation of the cortex* provokes "listening" movements of the ears in animals, and in man the perception of noises, but no motor response.
4. *Electrical potentials* (*action potentials*) have been registered at all levels of the auditory path and have been used for the localization of sound reception. Thus tones of a certain pitch give rise to potentials mainly in a definite part of the auditory cortex, which varies with the pitch. On the other hand, when there is localized destruction of certain places in the cochlea, electrical response to sound ceases in circumscribed areas of the auditory cortex. Maps have been drawn from data obtained by this type of experiment, giving the locus of perception of tones of different pitch in the temporal cortex of the dog and cat.[1] The results are similar in both species.

[1] Ades, H., F. A. Mettler, and E. Culler, *Am. J. Physiol.*, **125**, 15, 1939.

Summary. The auditory path arises in the cochlea. It is a four-neuron chain, ending on the temporal cortex of both sides. An outstanding feature of the auditory system is tonal systematization, evident in the cochlea, the auditory nerve, the medial geniculate body, and the temporal cortex, as if each tone provokes impulses that travel along a certain path conditioned by the pitch of the tone. Each cochlea sends impulses along the ipsilateral and contralateral paths, and action potentials can be registered on both sides when the ear on one side only is stimulated. Destruction of only one cochlea is followed by a very small loss of hearing. Unilateral destruction of the auditory area of the temporal cortex is well compensated by the remaining cortex; bilateral destruction is followed by a considerable reduction, but not the total suppression of hearing. The auditory area has no motor functions. The auditory area on the left side is necessary for the integration of the spoken word; its destruction produces a certain type of aphasia (Chap. 80).

SUMMARY OF THE MECHANISM OF AUDITION

Sounds, *i.e.*, vibrations of elastic bodies, are transmitted to the air. They cause variations in pressure in the external auditory meatus, which provoke vibrations in the tympanic membrane. This membrane reproduces sound waves and transmits them to the ossicle chain, which, acting as a lever, amplifies them and transmits them to the base of the stapes. Contraction of the muscles in the middle ear diminishes

[1] Woolsey, C., and E. Walzl, *Bull. Johns Hopkins Hosp.*, **71**, 344, 1942; Tanturi, *Am. J. Physiol.*, **141**, 397, 1944.

sensitiveness to tones of low pitch and protects the ear.

Movements of the base of the stapes are transmitted as pressure changes in the perilymph of the inner ear. These pressure changes, owing to the inextensibility of the osseous labyrinth, travel up the scala vestibuli and down the scala tympani to the membrane of the round window, which bulges out on receiving them. High tones cross from one scala to another near the base of the cochlea; low tones cross over near the helicotrema. On crossing, the pressure wave passes through the scala media and causes the basilar membrane to vibrate, at a definite place in relation to the pitch of the tone. This is the fundamental phenomenon in the recognition of pitch.

The vibration of the basilar membrane gives rise to cochlear microphonics, which reproduce the form and frequency of the sound waves. Their origin is still unknown, but they have been very useful in the study of audition.

The cochlea stimulates the auditory nerve. Tones of frequencies up to 900 cycles provoke responses in all the nerve fibers that are stimulated by tones of the pitch in question. The frequency of the impulses is the same as that of the sound waves. The ratio of tone frequency to impulse frequency is 1:1. When the frequency of the tone reaches 1,800, this ratio is 2:1; half the fibers are stimulated alternately. When the frequency of the tone reaches 2,700, the frequency of response again drops, and the ratio of sound frequency to impulse frequency is 3:1; one-third of the fibers are stimulated in rotation. When the tone frequency is 3,500 or more, there is no synchronization between sound waves and frequency of impulses.

The pitch of a sound depends on the part of the basilar membrane and the organ of Corti that vibrates and stimulates the corresponding fibers of the auditory nerve. There is localized projection of the cochlea on the medial geniculate body and the superior temporal gyrus, so that a tone of given frequency stimulates a definite place in the cochlea and is transmitted along a definite path to a definite place on the auditory cortex.

The intensity of a sound depends on the length of the basilar membrane that is set in vibration and therefore on the number of nerve fibers stimulated, although as yet a precise numerical relation between intensity of sound and number of fibers stimulated has not been established.

Masking is due to overlapping of pitches in the basilar membrane and to the competition for transmission of impulses along the same nerve fibers.

Stimulation of one ear provokes action potentials in the auditory paths and auditory cortex of both sides of the brain. Each cochlea, therefore, has bilateral cortical representation. This explains why loss of one ear, or unilateral destruction of the auditory paths or cortex, provokes little loss of hearing.

DISTURBANCES CAUSED BY SOUNDS. ULTRASONICS OR SUPERSONICS

Very loud sounds or noises, brief or prolonged, cause damage to the middle ear and cochlea, which in some cases is irreparable. This is the cause of a certain type of deafness in boiler workers, aviators (who always have faulty bone transmission after 3,000 hr. of flight), tank drivers, subjects exposed to explosions, etc.[1] In all these cases protective measures are taken in order to prevent deafness, *e.g.*, covering the ear or the external meatus, etc. Moreover, noise causes general disturbances even in normal subjects. There is difficulty in concentrating and sustaining attention, loss of sleep, increase in oxygen consumption (up to 25 per cent), increase in the work of the heart and respiratory organs, etc. Noise may cause a loss of efficiency up to 50 or even 60 per cent in factory workers. Convulsions and coma have been produced in rats by very loud sounds and by supersonics.

Supersonics are frequencies above 20,000 cycles; this is a conventional figure, because many subjects, especially old people, cannot hear sounds of more than 16,000 c.p.s., or even 12,000 c.p.s. Moreover, some animals hear supersonic vibrations inaudible to man. Subsonics are vibrations below 25 c.p.s., *i.e.*, below the minimum audible frequency.

Electrical energy can be converted into sound vibrations of very short wavelength, corresponding to frequencies between 20,000 and 500,000,000 c.p.s. In one kind of apparatus an iron bar or tube is placed in a magnetic field so that it

[1] Armstrong, H., and J. Heim, *J. A. M. A.*, **109**, 417, 1937.

lengthens and shortens very rapidly; powerful supersonic vibrations are emitted from both ends. Piezoelectric crystals are also used and can give very high sound vibrations.

Supersonic vibrations transport energy. They travel in a straight line, but they can be reflected and therefore directed along a given path. By measuring the time taken by supersonic waves to travel to and from a surface on which they are reflected, it is possible to calculate the distance to the reflecting object. In clinical medicine this has been used for the location of brain tumors, gallstones, etc.[1]

The results of the experimental application of supersonics to animals vary with the animal species and with the duration and frequency of the vibration. The energy transmitted increases with the frequency. Many bacteria and other unicellular organisms are not disturbed by these very short sound waves, because they ride on them; other cells, such as erythrocytes, may be severely damaged and explode. Fishes, frogs, and other animals have been killed by supersonics. Supersonics have been applied in human therapeutics, but even in "therapeutic" doses, they have been shown to produce lesions in the bones, liver, and nerve centers of experimental animals.[2]

The mechanism by which the damage is caused varies. The most important is "cavitation," *i.e.*, the production of submicroscopic bubbles in tissue fluids (owing to the changes in pressure), which disorganize tissue structure. The process is instantaneous; *e.g.*, eggs of *Arbacia* are destroyed in 1/1200 sec., as can be demonstrated by photographic registration. Supersonics also have chemical activity, which is evident, for example, in the depolymerization of starch, gelatin, etc. Physical effects are produced, *e.g.*, sulfathiazol crystals are reduced to microscopic dimensions. They have thermal effects, causing a rise in temperature in solutions and emulsions, in tissues and in the whole animal, *e.g.*, they can raise the body temperature of mice to 45°C. They have other physical and chemical effects, such as the formation of microcrystals (sulfathiazol, etc.); mixtures of nonmixable fluids, *e.g.*, oil and water; depolymerization of starch and gelatine; etc.

[1] This method is used for measuring the depth of the sea and for locating submarines, shoals of fish, etc.

[2] COUNCIL ON PHYSICAL MEDICINE AND REHABILITATION, *J. A. M. A.*, **148**, 646, 1952.

BIBLIOGRAPHY

BEATTY, R., "Hearing in Men and Animals," G. Bell, London, 1932.

DAVIS, H., and S. BARTLEY, Sense Organs, *Ann. Rev. Physiol.*, **5**, 345, 1943.

FLETCHER, H., "Speech and Hearing," Macmillan, London, 1929.

GLASSER, O., "Medical Physics," Year Book Publishers, Inc., Chicago, 1950.

GOLDSMITH, M., Ultrasounds at the Service of Man, *Bol. Cent. Coop. Cient. Unesco*, **1**, 9, 1952.

GREGG, E. C., Ultrasonics. Biological Effects, in O. Glasser's "Medical Physics," Vol. 2, Year Book Publishers, Inc., Chicago, 1950, p. 1133.

HARTLINE, H. K., Sense Organs, *Ann. Rev. Biochem.*, **4**, 445, 1942.

HELMHOLTZ, H. L., "Sensations of Tone," Longmans, New York, 1930.

KERRIDGE, P. M., Recent Advances in Knowledge Concerning Hearing and Speech, *Physiol. Rev.*, **18**, 59, 1938.

MUNSON, W. H., Sound and Hearing, in O. Glasser's "Medical Physics," Vol. 2, Year Book Publishers, Inc., Chicago, 1950, p. 995.

RANKE, A., Physiologie der Schnecke und des Cortischen Organs, *Ergebn. d. Physiol.*, **37**, 12, 1935.

RICHARDSON, J. R., and E. M. HOLMES, Anatomy and Physiology of the Ear, *Arch. Otolaryng.*, **35**, 480, 1942.

STEVENS, S., and H. DAVIS, "Hearing: Its Psychology and Physiology," Wiley, New York, 1938.

STUHLMAN, O., Hearing. Mechanism, in O. Glasser's "Medical Physics," Vol. 2, Year Book Publishers, Inc., Chicago, 1950, p. 385.

WEVER, E. G., "Theory of Hearing," Wiley, New York, 1949

CHAPTER 80

Voice and Speech

Language is the means by which animals and man communicate and express feeling and thought.

There are two kinds of language: (*a*) spontaneous, natural, or primitive language; (*b*) conventional language, in which certain signs are given a definite meaning. The latter must be learned in order to be understood and used.

Natural or primitive language. This type of language is used by all the higher animals, including man; it becomes more developed as the inteligence of the animal increases. States of mind are expressed in many ways, which can be classified into three groups: (*a*) attitudes, which indicate feeling or mood, such as the extended limbs, arched back, erect fur, and lashing tail of an angry cat; (*b*) gestures, such as the pricking of the ears and movements of the head in the act of listening; (*c*) emission of sounds. In vertebrates, sounds are produced in the larynx and modified by changes in the shape of accessory structures (throat, mouth, and nasal cavities), thus forming the voice. Each animal species has its own voice, which changes in the expression of different emotional states; *e.g.*, it is easy to recognize the differences in the clucking of a hen calling her chicks, her cackling after she has laid an egg, or the screeches she emits when frightened.

Conventional language. This type of language is used exclusively by man, who reinforces its meaning by the simultaneous use of primitive language. There are two forms of conventional language, the spoken and the written. In the former the voice is emitted in a series of articulate elementary sounds or phones, known as vowels and consonants, which are combined into words. Words are symbols for ideas, and they form the basis of speech. In written language, graphic symbols represent ideas (ideographic writing) or sounds (phonetic writing). In the latter the symbols for the different vowels or consonants are combined into words as in spoken language.

Speech, *i.e.*, the expression of thought by articulate sounds, is the most highly developed form of language. A series of complicated mechanisms is necessary for speech. In the formation of the spoken word, by means of the auditory apparatus (ear and nerve centers), certain articulate sounds are perceived, distinguished from others, and endowed with meaning, *i.e.*, associated with certain visual, tactile, or other sensations, and recognized as symbols for these. These sounds are reproduced by adequate movements of the respiratory muscles, the larynx, and the resonating apparatus, coordinated by special nerve centers. Hearing exerts a predominant influence on speech, not only in childhood when speech is learned but throughout life, as is evident by the serious disturbances in speech caused by deafness. Deprivation or serious deficiency of hearing in early childhood causes mutism (deaf-mutism).

In the following paragraphs the different physiological aspects of speech will be considered: (*a*) voice production and the characteristics of the human voice; (*b*) the acoustics of speech; (*c*) the nervous mechanism of speech; (*d*) deaf-mutism.

VOICE PRODUCTION

There is great similarity between the apparatus for the production of the voice and musical instruments. In both there are (*a*) a force that puts a vibrating mechanism into action; (*b*) the vibrating mechanism itself; (*c*) a resonator that reinforces certain vibrations. In the piano these are respectively the hands of the pianist, the strings, and the sounding board of

the instrument. In a wind instrument they are the blast emitted by the player's lungs (the bellows in the organ), the reed (*e.g.*, in the clarinet) or the player's lips (*e.g.*, in the trumpet), and the tube of the instrument. In the instrument that produces the human voice, the force is given by the blast sent through the trachea by the lungs, which act as a bellows; the true vocal cords form the vibrating mechanism; and the resonator is made up of all the supraglottic cavities, *i.e.*, the upper part of the larynx, the pharynx, the mouth, and the nose.

The Respiratory Bellows

The voice is produced by the blast passing through the glottis. Immediately before speaking, the glottis is closed by the approximation of the vocal cords. The abdominal and thoracic muscles compress the lungs, increasing the pressure in the respiratory tract. When the pressure has reached a certain level, the glottis opens and the voice is emitted. The primary importance of the rise in pressure is demonstrated as follows: (*a*) if the trachea is opened, there is no rise in pressure and no voice; (*b*) if intratracheal pressure is measured in a tracheotomized subject (an experiment performed by Cagniard de Latour for the first time), it will be seen to rise to 16 cm. H_2O in ordinary conversation, to 30 cm. when the subject plays the clarinet, and to 94 cm. when he emits shrill cries.

Thoracic and abdominal muscles contribute to this increase in pressure in both sexes. This can be demonstrated by recording the action potentials of the different muscles. Thoracic muscles play the most important part in men and women, especially in the latter, and their activity increases in rapid talking.

The *intensity* or *loudness* of a voice is proportional to the amplitude of the vibrations of the vocal cords, which increase with the force of the expiratory blast. This can be demonstrated on an isolated larynx, varying the pressure of the air which is blown through it. Loudness of voice is measured in decibels. There is a difference of approximately 100 db. between the softest whisper and the loudest cry. In ordinary conversation, loudness of voice varies between 40 and 50 db.

The Vibrating Mechanism. The Larynx

The larynx is a short tube situated between the trachea and the pharynx. It can be considered as the modified upper part of the trachea.

ANATOMY

The larynx is formed by a framework of four cartilages, joined by fibrous tissue and ligaments, and several small muscles, which play a part in the production of voice. The lumen of the larynx is narrowed in three places; the lowest one, formed by the true vocal cords, is an essential part of the mechanism of voice production. The larynx is lined by a ciliated epithelium, except over the inner margin of the true vocal cords, which is covered by a stratified epithelium. A man's larynx is about one-third larger than a woman's, which is about one-third larger than that of a child at birth. The larynx grows rapidly up to the age of three years; at puberty there is another period of rapid growth, especially in boys.

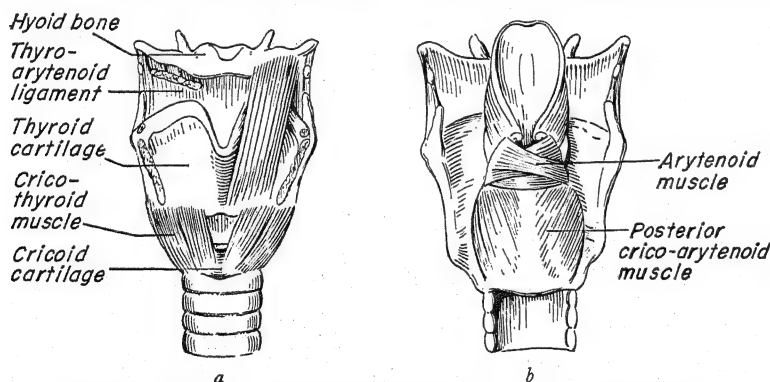



Fig. 461. Muscles of the human larynx. *a*, front view; *b*, back view.

Cartilages. The cricoid cartilage is the lowest. It is a modified tracheal ring shaped like a

signet ring, with the seal plate in the posterior aspect. The upper margin of this plate articulates with the base of the two pyramid-shaped arytenoid cartilages. These cartilages can rotate on their base around a vertical axis, and the whole cartilages can slide outward or inward so that their inner margins draw apart or come together. The cartilages have two processes: a median, or vocal, process on which the true vocal cord is inserted, and a lateral, or muscular, process on which several muscles are inserted (Fig. 461).

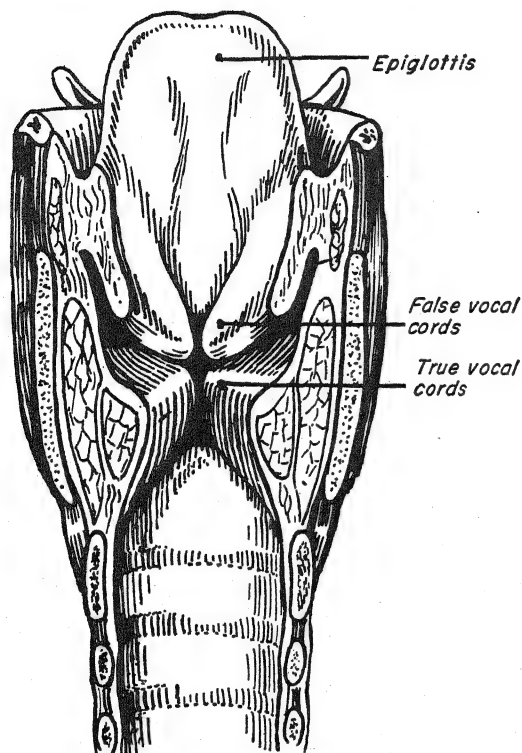



FIG. 462. Transverse section of human larynx. The upper surface of the true vocal cords is flat; their inner margin is sharp and the lower surface is concave toward the mid-line (*conus elasticus*). Both vocal cords form an arch at the summit of the trachea, which resists the expiratory blast when the glottis is closed.

The thyroid cartilage forms the anterior and lateral walls of the larynx. It is placed under the skin of the neck, forming the protuberance known as "Adam's apple" in the male. It consists of two vertical plates joined in the mid-line by fibrous tissue. The plates form an angle with its vertex directed forward. The anterior extremities of the true vocal cords are inserted on this cartilage. Ligaments and fibrous membranes join the thyroid cartilage to the hyoid bone above and the cricoid cartilage below.

The true vocal cords. The glottis. The true vocal cords are inserted in the angle formed by the thyroid cartilage. They are directed horizontally and backward, ending by insertion on the median or vocal process of the arytenoid cartilage. In a frontal section they appear as triangular structures with an upper horizontal surface and a lower median concave surface; the inner margin is formed by the union of these two surfaces, and in certain cases comes into contact with the inner margin of the vocal cord on the opposite side (Fig. 462). They are formed by the inner fibers of the thyroarytenoid muscle and by a fibrous ligament. The mucosa covering the vocal cords is made up of cylindrical ciliated cells, with mucous glands, except over the inner margin, where there is a stratified epithelium. These glands continuously secrete mucus, which lubricates the vocal cords. There are always a few drops of mucus on the inner margin of the cords. Between the inner margins of the two vocal cords there is an angular space, with the vertex directed forward, which is continuous with the interarytenoid space. Seen from above, this space has the shape of a lance head. The anterior part between the vocal cords is known as the vocal glottis. Its length is approximately 18 mm. in men and 13 mm. in women; the posterior part is the cartilaginous glottis, which is 7 to 8 mm. in length.

Above and parallel to the vocal cords are two mucous ridges known as the false vocal cords. On each side, between the false vocal cords and the lateral wall of the larynx, there is a recess called the ventricle of Morgagni. The mucosa at this level covers loose adenoid tissue, which can be easily infiltrated by fluid, becoming edematous. This edema causes a serious obstacle to respiration, and may even cause death by asphyxia.

The *air sacs* are rudimentary in man, merely a depression in the ventricle of Morgagni. In certain animals they are large sacs full of air. They open into the mouth in frogs, the larynx in the anthropoids, the trachea in ducks, the lungs in chickens. They act as air reservoirs, used in a struggle when the breath is held (they swell out in anthropoids when angered), or to keep up the air blast during singing in birds. In pathologic cases in man, they can be abnormally large (laryngocele).

Muscles and movements of the vocal cords. All the muscles of the larynx except the arytenoid muscles are paired and symmetrical. All of them act on the vocal cords, by approximating them, separating them, or changing their tension.

FIG. 463. Diagram of horizontal section of the larynx at the level of the arytenoid cartilages. I, contraction of the posterior cricoarytenoid muscle, rotation of the arytenoid cartilages in the direction of the arrows, and opening of the glottis; II, contraction of the lateral cricoarytenoid muscles, rotation of the arytenoid cartilages in the direction of the arrows, and first stage in closure of the glottis; III, contraction of the arytenoid muscles and approximation of arytenoid cartilages; IV, contraction of the thyroarytenoid muscles, shortening and complete closure of the glottis.

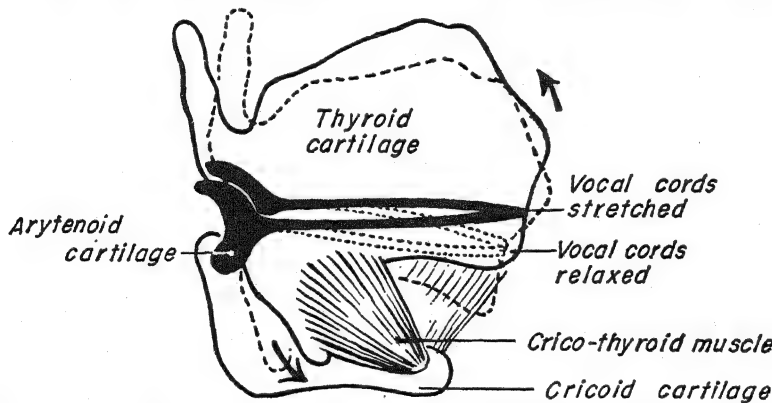



FIG. 464. Contraction of the cricothyroid muscles. The vocal cords are stretched and drawn together, closing the glottis, when the posterior part of the thyroid cartilage is pulled down. (*After Pressman.*)

1. The vocal cords are separated and the glottis is dilated by the posterior cricoarytenoid muscles, which rotate the arytenoid cartilages (Fig. 463, I).
2. The vocal cords are approximated and the glottis is constricted by three groups of muscles: the lateral cricoarytenoid muscles, which rotate the arytenoid cartilages around a vertical axis in the opposite direction to the rotation produced by the posterior cricoarytenoid muscles; the arytenoid muscles, which draw the arytenoid cartilages together; and the lateral fibers of the thyroarytenoid muscles, inserted on the muscular process of the arytenoid cartilage, which rotate these cartilages and draw them forward, completing the closure of the glottis (Fig. 463, II, III, and IV).
3. The tension of the vocal cords is increased by two pairs of muscles: the cricothyroid muscles cause the anterior angle of the thyroid cartilage to move upward and forward, thus stretching the vocal cords by separating their anterior and posterior insertions (Fig. 464); and the inner fibers of the thyroarytenoid muscle, which forms the musculus vocalis in the vocal cord. The fibers of this muscle are inserted on the thyroid or arytenoid cartilages by one end and on the vocal cord by the other. They contract in separate bundles; they sharpen the margin of the vocal cords, diminish their surface of contact, and modify the contour of the glottis. They are fast fibers, which respond very rapidly to stimulation.

Nerves. The upper and lower laryngeal nerves are branches of the vagus nerve. The upper laryngeal nerve sends a lateral branch to the cricothyroid muscle and has an inner branch which gives out sensory fibers to the whole larynx. There are also vasomotor and secretory fibers to the mucous glands. Section of this nerve anesthetizes the larynx and suppresses reflex coughing when a foreign body drops into the larynx. It also causes paralysis of the cricothyroid muscle, and the voice becomes hoarse because of the lower tension of the vocal cord.

The recurrent laryngeal nerves send out motor fibers to all the muscles of the larynx except the cricothyroid muscles. Section of one lower laryngeal paralyzes the homolateral vocal cord, which then approximates the mid-line. The voice becomes bitonal, and the vocal cord is not separated on inspiration. Bilateral section of the recurrent laryngeals suppresses phonation (aphonia) and causes serious disturbances in respiration; the glottis is not dilated in inspiration, and it may even be closed by further approximation of the vocal cords (*e.g.*, during muscular exercise). In young animals death by asphyxia may occur because the cartilaginous glottis may also be closed by approximation of the arytenoid cartilages; in the adult these cartilages are hardened, and the posterior part of the glottis remains open. In normal conditions the dilators of the glottis predominate over the constrictors, and the glottis has a width of 12 to 14 mm. After section of the four laryngeal nerves, the width of the glottis is reduced to 4 or 5 mm.

PHYSIOLOGY

The main function of the larynx in all animals is to assure that only air passes into the lungs. In man and other species a sound is produced by the vibrations of the vocal cords, which is modified in the supraglottic resonators and converted into the voice. The laryngeal sound can be emitted with different loudness and pitch, because it can be modified by the following factors: (*a*) the strength of the blast coming from the lungs; (*b*) the size and shape of the glottis; (*c*) the tension of the vocal cords.

Functions of the larynx in voice production. The laryngoscope. Müller in 1839 showed that a vocal sound was produced by the isolated larynx when a blast was sent into it from a bellows through the trachea. He also demonstrated the changes in loudness and pitch produced by varying the strength of the blast and the tension of the vocal cords, which were stretched more or less by weights pulling on the thyroid cartilage.

Magendie pointed out that the vocal tone is produced at the level of the true vocal cords. He separated the larynx of dogs by a transverse section below the hyoid bone and saw that when the animal barked the glottis was constricted. Extirpation of the vocal cords in animals leaves them voiceless. Models of the larynx can be made which show how the voice sound changes when the tension of the vocal cords is modified.

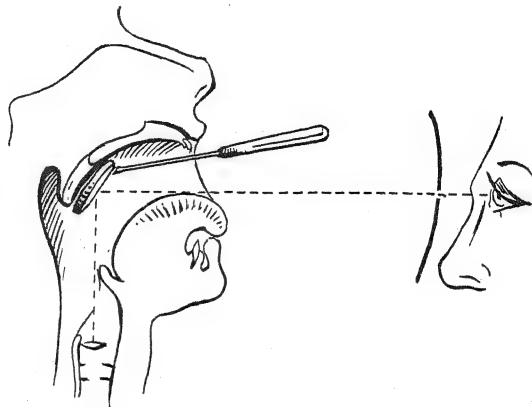



FIG. 465. Observation of the larynx by means of the laryngoscope.

Manuel Garcia, a tenor and singing teacher, in 1855 invented the laryngoscope, which is still used by laryngologists. It consists of two mirrors. One is a large concave mirror with a central aperture placed in front of the observer's eye. It reflects a beam of light onto a second small mirror fixed at an angle on a handle so that it can be placed close to the soft palate of the patient, who holds his mouth wide open. The observer can thus see the image of the larynx reflected in the small mirror (Fig. 465).

Periscopes introduced into the mouth or through the nose have also been used to visualize the larynx. The photographic method was first used by French in 1886. A series of pictures (plain or colored) are taken on fast films and projected with the slow-motion camera or by means of a stroboscope.[1] The images are associated with the production of vocal tones. No two larynges are exactly alike, but they all have certain common characteristics which make it

[1] An instrument for showing a series of pictures rapidly, so that an illusion of motion is produced.

possible to give a general description of laryngeal movements in voice production.

When the larynx passes from the resting position to that necessary for emission of sound, the first movement is the approximation of the vocal cords until they come into contact with each other; this is known as the phonation position of the cords. The glottis is thus closed, and air pressure in the trachea increases. Approximation and tension of the vocal cords are brought about by contraction of the laryngeal muscles as described above. The countertension mechanism formed by the internal fibers of the thyroarytenoid muscle then enters into activity; the cords are separated, in part owing to retraction of their internal margin; and the contour of the glottis changes. Finally, as the voice is emitted the internal margins of the cords enter into fine vibratory movements; there is periodic expansion and narrowing of the glottis.

THE PHYSICAL MECHANISM OF VOICE PRODUCTION

The laryngeal sound is produced by the passage of air under pressure through the narrow glottis. Vibrations of the vocal cords interrupt the passage of air, which becomes intermittent. Thus eddies are formed in the supraglottic cavities.

There is still some discussion as to whether the blasts of air that pass through the glottis act directly on the air in the supraglottic cavities, causing it to vibrate and making the vocal cords vibrate in response, or whether the cords are set in vibration by the blast of air passing through the glottis, as wind causes vibration in stretched wires. If the cords were set in vibration directly by the blast of air passing through them, they would be submitted to very high tension in the emission of certain tones, higher than they could resist. The other hypothesis, in which the larynx has been compared to several wind instruments with or without vibrating reeds, is the more probable. So far, however, no fully satisfactory explanation of the mechanism that sets the cords in vibration has been given.

FIG. 466. Tonal changes in the human voice from birth to puberty. Solid line, boys; broken line, girls. (*According to data given by Paulsen and by Gutzmann.*)

Pitch. The "highness" or "lowness" of a sound depends on the cycles per second of its vibrations. The number of tones which can be emitted by the human voice and their position in the musical scale vary with age and sex.

The frequency range of the newborn is limited to approximately three semitones. As the child grows the frequency range increases mainly by the addition of higher tones, but also of a few low tones (Fig. 466). The highest frequency is attained usually around the eleventh year; at this age girls can emit very high sounds, seldom reached by sopranos in later years. At the age of puberty there is little difference between the voices of boys and girls, except that the latter can usually reach a few notes higher.

Structural modifications in the larynx which occur at puberty cause the voice to change suddenly. The pitch is lowered one octave[1] in boys and only two notes in girls. Thus the voices of boys change from soprano to tenor and from contralto to bass.

as has been shown by observations in laryngectomized subjects and by observations with the laryngoscope.

Effects of laryngectomy. Laryngectomized subjects can recover speech by learning how to swallow air, how to retain it in the esophagus by constriction of the pharyngoesophageal sphincter, and then how to control its passage into the pharynx and mouth. Words can be formed, but there are no tonal variations and it is not possible to sing.

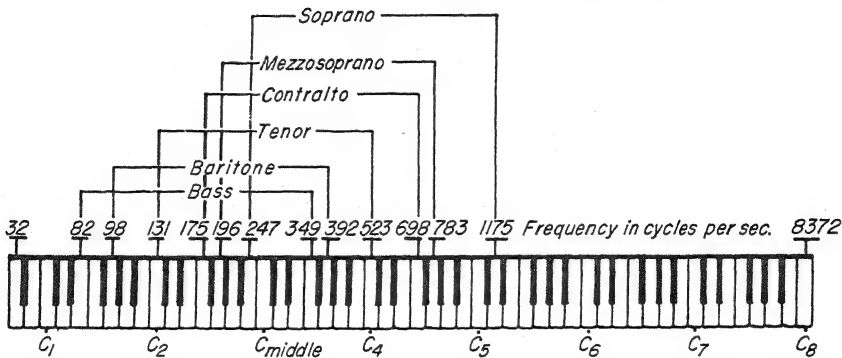



Fig. 467. Ranges of human voice in the musical scale compared with the keyboard of a piano.

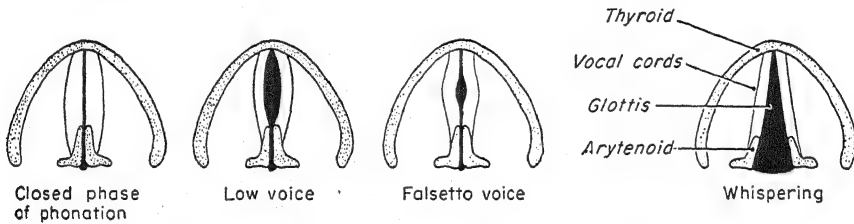



Fig. 468. Shape of the glottis during the emission of the voice.

The voices of adult men and women differ in their position in the musical scale. Three voices are distinguished in each sex: bass, baritone, and tenor in men; contralto, mezzo-soprano, and soprano in women (Fig. 467). The pitch of the voice is often lowered with age; thus a tenor may change into a baritone and a soprano into a mezzo-soprano. The pitch range in ordinary conversation is usually not more than half an octave. It increases in recitation and goes up to two octaves in singing. The whole range of human voices covers four octaves; two and one-fifth octaves for each sex, with an overlap of one-fifth octave in the middle of the scale.

The voice does not change at puberty in eunuchs. It retains its infantile characteristics, but pitch range and intensity increase. In the past, famous choirs were made up of eunuch voices.

Changes in pitch. Voice registers. Tonal changes in the voice are produced in the larynx,

[1] The frequencies of vibration of two notes separated by one octave are as 1:2.

Observation of the larynx during voice production. According to the pitch of the sound emitted, the larynx moves up or down; the vocal cords change in length, position, and tension, and the shape of the glottis varies (Figs. 468 and 469). The force of the blast also varies.

In the production of *low tones* the larynx rises a little, the cords lengthen and relax, they come into contact with each other, and the whole cord vibrates. The glottis is open, especially in its posterior part, but not so much as during breathing.

During the emission of *high tones* the larynx rises considerably, the cords shorten and become tense, their margin sharpens and is the only part that vibrates. The glottis is reduced to a narrow slit. In the production of very high tones, such as those of the head register or falsetto, a special

damping mechanism comes into play. This mechanism has been compared by Pressman to the effect produced on the string of a violin by pressure exerted with a finger; the string is divided into two parts, only one of which vibrates. The vocal cords are brought together and touch, except on a small length of their anterior end; the smaller the length of this free vibrating part, the higher the tone emitted (Table 112).

Table 112. Position of the True Vocal Cords and Shape of Glottis and Supraglottic Space in the Emission of Low and High Sounds

	Sound		
	Low	*High*	*Very high (falsetto)*
Vocal cords			
Length	Short	Long	Long
Tension	Low	High	High
Vibrations	Total	Inner margin	Inner margin of anterior part
Glottis			
Position	Low	High	High
Shape	Racket, handle backward	Lineal	Lineal, posterior part closed
Supraglottic space	Lengthened	Shortened	Much shortened
Subjective vibrations	Chest	Head	Head

In whispering the arytenoid cartilages and the vocal cords are widely separated. The cords vibrate irregularly. The resonators modify the sound thus produced, and it becomes articulated into words but remains toneless.

The effect of the expiratory blast. Pitch rises simultaneously with loudness, as the force of the blast increases. This phenomenon can be easily demonstrated in an isolated larynx, and it explains the difficulty in singing softly tones of high pitch.

In *summary*, the pitch of the voice depends on the length and tension of the vocal cords, the shape of the glottis, and the force of the blast. If the vocal cords are suppressed, the voice is toneless.

Registers. The capacity to emit sounds of different pitch is outstanding in man. In other species it is much more restricted; thus cattle emit only low tones, and monkeys high tones. Singers pass smoothly from high to low tones and vice versa. In music, voice range is usually divided into registers. The low or chest register comprises low tones rich in harmonic overtones; it is given this name because vibrations are objectively and subjectively located in the chest. The head register comprises high tones, with few harmonic overtones; the vibrations are subjectively located in the head; hence its name.

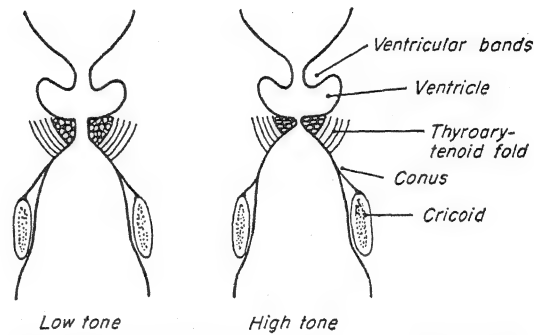



Fig. 469. Shape of air tube in the emission of low and high tones.

The false vocal cords. These structures are not necessary for the production of voice, and in some animals (*e.g.*, the cat) they do not exist. In very good singers they are separated and placed against the wall of the larynx, giving a free passage to the sound waves; in others they are brought near the mid-line and give a harsh quality to the voice. Their main effect is the production of overtones by entering into vibration in response to the vibrations of the glottis. When the true vocal cords are extirpated, the false ones may replace them to a certain extent, giving a harsh sound of low intensity and little tonal variation.

The Resonating Mechanism

All the space through which the expired air passes, *i.e.*, the lung, trachea, larynx, pharynx, mouth, and nasal cavities, are called the resonators, because of the part they play in voice production.

The resonators are not static, but dynamic; their shape changes, adapting itself to the pronunciation of vowels and consonants needed in the articulation of words. The resonators consist of static parts, such as the bony palate, the nasal canal, etc., and movable parts such as the soft palate and uvula, the tongue, lips, cheeks,

etc. Malformations of the resonators cause considerable disturbances in the quality of the voice. The effects of a harelip or cleft palate, the absence or incorrect implantation of teeth, and the presence of adenoids in the pharynx are well-known causes of defects in speech. Facial paralysis, or paralysis of the soft palate or the tongue, also causes speech disturbances.

The resonators increase the volume of the weak laryngeal sound, and they reinforce some of the overtones, giving the voice its individual timbre, which permits the recognition of each person's voice. Thus also is it possible to distinguish the voice of a tenor from that of a baritone even when both sing the same note. More important still is the part played by the resonators in forming the vowels and consonants which make up words.

Changes that take place in the resonators in the course of speech have been recorded photographically, by x-rays taken after making the different parts opaque with barium, and by cinema films. It has thus been seen that there are considerable individual variations in the position of the tongue on pronouncing the same voice sounds.

Mechanics of vowel formation. Vowels are formed by a continuous sound, easy to recognize, produced by the arrival of the laryngeal sound at the mouth and the reinforcement of some of its overtones. Helmholtz showed that each one is formed by special tones, and he was able to reproduce the vowel sounds by a system of tuning forks, each one giving a single pure tone. The laryngeal sound is not indispensable, since it can be replaced. For example, a vowel sound can be produced by placing the mouth in the adequate position for its pronunciation and percussing the cheek with a finger. It can also be produced by whispering, and laryngectomized persons can pronounce vowels correctly.

The resonators adopt a particular position for the pronunciation of each vowel sound. There are three typical or extreme vowel sounds: *ä* (as in "far"), *ē* (as in "eve"), and *ōō* (as in "tool"). Other vowel sounds derived from these vary from one language to another. The *ä* sound is considered as the fundamental vowel; it is pronounced with the tube open to a maximum, and it is easy to sing. For the pronunciation of *ē* and *ōō* the tube is narrowed, and these sounds are not so easy to sing. Changes in the resonating tube in the pronunciation of fundamental vowel sounds are summarized in Table 113.

Table 113. Pronunciation of Fundamental Vowel Sounds

	ä	*e*	*ōō*
Position of larynx	Intermediate	Very high	Very high
Opening of mouth	Wide open	Narrowed by tongue and palate	Narrowed by lips
Position of tongue	Retracted down	Close to palate	Tip down, root high
Number of paces at which vowel sounds can be heard	360	300	280

Mechanics of consonant formation. Consonants are sounds or noises produced in the mouth by suddenly placing an obstacle to the passage of the expiratory air blast. They are interrupted vowel sounds. The site where this obstacle is placed is known as the area of articulation. The laryngeal sound is not indispensable.

According to the place of formation, consonants are classified into

1. Labials or lip consonants, which are bilabial (*b, p, m*) when formed between the two lips, or labiodental (*v, f*) when formed between the lower lip and upper teeth, and labiogutturals (*wh* in when.)
2. Dentals, formed by the tip of the tongue extended between the edges of the front teeth (*th* in "then"), or by the tip of the tongue touching the back of the upper front teeth or gums (*d, t*), or by forming a narrow passage between the back of the tongue and the upper front teeth (*s, z*). In some cases the tip of the tongue is raised (*sh* in "shoe").
3. Palatals, in which the tongue is raised up against the palate (*y* in "yell," *r* in "round").
4. Gutturals, formed by raising the root of the tongue toward the soft palate (*g* in "go").

According to the manner of formation, the consonants are called (*a*) stopped consonants,

when the blast of expired air is completely stopped at some point in the air passage (*b*, *p*, *t*); (*b*) nasal consonants, when the mouth passage is completely stopped but the soft palate does not close the passage into the nasal cavities (*m*, *n*); (*c*) open consonants, when the blast is not stopped but is forced through a narrow passage (*s*, *z*, *th*, *ch*).

Summary

Speech is produced as follows:

1. Words are always formed during expiration. If the subject is inspiring, the movement is interrupted and replaced by expiration. At the same time the vocal cords adopt the phonation position, becoming tense and closing the glottis, by contraction of the appropriate laryngeal muscles. Intratracheal pressure rises owing to the contraction of expiratory muscles in the thorax and abdomen.
2. The glottis is partially opened. The internal fibers of the thyroarytenoid muscle (musculus vocalis) contract to give the glottis the shape adequate for the tone to be emitted. Pitch is due to the tension and length of the vocal cords, to the shape of the glottis, and to a certain extent to the force of the blast, which conditions the loudness of the voice.
3. The air escapes through the glottis under pressure. The vocal cords are separated and vibrate. The air passing in blasts into the supraglottic space also vibrates.
4. The laryngeal sound (vocal tone) thus produced is modified in the resonators, which change in shape so that some of the overtones are reinforced and the passage of the expiratory blast is checked. Thus vowels and consonants are formed, which are combined into words, and the voice acquires its characteristic quality or timbre.

THE ACOUSTICS OF SPEECH

The physical analysis of speech was begun a century ago. Helmholtz and König were among the first to make important contributions to this subject, but progress was relatively slow until the advent of electromechanical devices for recording, producing, and reproducing sound. Advancement along these lines has proceeded simultaneously with the development of the phonograph, telephone, radio, etc.

Methods. The majority of the methods that have been used are not sufficiently sensitive. Resonators of different size and periodicity were first used to determine the components of a sound. König added to the resonators a "sound flame," which oscillates with the vibrations of the resonator and can be photographed so as to obtain a permanent record. Sensitive membranes, which vibrate on the impact of the voice, have also been used. Phonographic registration of the voice and analysis of the records have recently been considerably developed. More sensitive and accurate methods are based on the conversion of sound waves into electric potentials; these electroacoustic methods are now commonly used in the registration of the sounds and noises of speech and give records that are easy to analyze. The oscillograph registers on a fluoroscope, or on paper, wave sounds which have been taken up by a microphone, converted into an electric current, and amplified. The wave recorded is composite, made up by the different tones which constitute the sound registered, so it must be submitted to a rather laborious analysis. In the spectrographic method the sound is first recorded on a gramophone record, then reproduced, and by using adequate sound filters each tone is recorded. This mechanism is similar to that of the inner ear. Vowel sounds thus analyzed have been resynthesized.

Physical Properties of Fundamental Voice Sounds (Phones)

According to Gemelli and Pastori, a phone is "a speech sound or noise which cannot be divided into simpler elements of speech, but only into simple oscillatory phenomena." Phones are therefore phonetic units, *i.e.*, vowels and consonants.

Vowels. Vowel sounds have been analyzed as they are pronounced within a word (real vowels) or separately (vowel types). A vowel type gives a curve with an initial phase, a central or uniform phase, and a final phase. In real vowels (pronounced within a word) the first and last phases are transitional between the preceding and following sound. Whispering and singing give very clear curves.

Each vowel has a typical curve that can be easily recognized. This constant element has a variable part added which depends on the tone of voice, the type of pronunciation, etc., *i.e.*, the individual characteristics of the phonetic appa-

ratus of the subject, and regional peculiarities in pronunciation. The fundamental nature of the curve can nevertheless always be recognized.

Analysis of the components of the curve of each vowel sound, by Fourier's mathematical analysis or Dietsch's photoelectric method, has shown that all the vowels are formed by algebraic summation of two or more simple tones of different rhythm and form. These tones are typical for each of the vowel sounds, some of which (*e.g.*, ŏ and o͞o) can be reproduced by summation of the simple components. The most complex curve is the one corresponding to ä, which is the most highly developed vowel sound; ē and o͞o are very simple (Fig. 470). Other vowel sounds are of intermediate complexity.

Fig. 470. Oscillographic registration of sound waves corresponding to pure vowels (upper curve). Below, the different components of the complex wave obtained by Fourier's analysis. The ä sound corresponds to a male voice; it has a fundamental frequency of 235 c.p.s. The o͞o and ē sounds correspond to a female voice; their frequencies are 430 and 586 respectively. (*After Gemelli and Pastori.*)

Consonants. Consonants also have typical, recognizable curves. These curves have been analyzed into their components, and some of the consonant sounds have been reproduced by synthesis, but the results have not been as perfect as with vowel sounds.

Some of the consonants are sounds, *i.e.*, they are produced by regular cycles; others are due to irregular oscillations, *i.e.*, they are noises. The former are called voiced consonants, or sonants, or semivowels (*l*, *n*, *m*, *v*); the latter are true consonants, or voiceless breath consonants.

Words. Oscillographic and spectrographic methods have been applied to the study of words and sentences. The spectrogram[1] shows, on the fluoroscope or on paper, simultaneously all the different frequencies and their changes so that a complete analysis of the word can be made. This method which makes words "visible" has been used in the instruction of deaf-mutes and in voice training. Reading of the records must, of course, be learned first.

Some applications of acoustics. The physical analysis of speech is of great practical importance in the construction of auditoriums with good "acoustics" and in improving audibility in telephonic communication or faithfulness in the reproduction of sound by phonographs and radio sets. For example, a vowel sound must last a minimum time of 7.7 msec. (*i.e.*, 2 cycles); if it is below this minimum, it will not be heard over the telephone or the radio. The optimum time for perception of a phone is 0.1 to 0.033 sec.; if it lasts more than 0.5 sec. it is not heard well. In an auditorium sounds are heard by direct transmission from the source, but mainly (90 per cent) indirectly after they have been

[1] Potter, R. K., Sensory Devices. Sound Portrayal, in O. Glasser's "Medical Physics," Vol. 2, Year Book Publishers, Chicago, 1950, p. 979.

echoed from the walls and other surfaces. The ear receives these reflected waves with a delay that increases with the distance of the reflected surface (Fig. 471). Reverberation (as this phenomenon is called) should not prolong the duration of the sound for more than 0.5 sec.

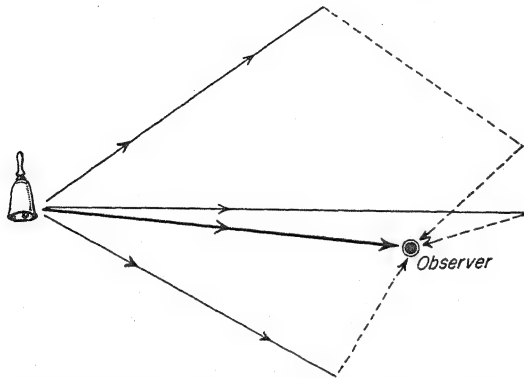



FIG. 471. Reverberation. The sound of the bell reaches the observer directly (thick arrow) and indirectly (thin arrows) after reflection on the walls and objects (dotted arrows), with a delay proportional to the length of the path traveled by the reflected sound waves. The stimulus is thus prolonged and the sound continues to be heard after the bell has ceased ringing.

The intensity of the sound is conditioned by the fundamental tone; the overtones condition the possibility of understanding its significance. Crandall eliminated by adequate filtration all waves of frequencies below 500 c.p.s. from certain speech sounds; the intensity diminished 60 per cent, but audibility diminished only 2 per cent. Therefore the high tones in the voice, *i.e.*, those of 1,000 to 2,000 c.p.s., which carry the overtones, are the important ones for audibility and are those which should be reinforced in order to improve hearing.

THE NERVOUS MECHANISM OF SPEECH

The clinical observations of Broca (1861) and Wernicke (1874) awakened interest in the study of the nervous mechanism of speech. A diagrammatic conception of speech formation was built up by artificially separating elementary components of speech and assigning to each a cortical center. In the first quarter of this century, further clinical observations led to the critical revision of this diagrammatic interpretation of the neural mechanism of speech, and new concepts were gradually developed (Pierre Marie, 1906; Head, 1917; Goldstein, 1927).

The learning of speech requires the development of two closely associated processes, reception and understanding on the one hand and expression on the other. The first process consists of two parts: (*a*) stimulation of the auditory receptors and centers; (*b*) the passage of impulses from these centers to mechanisms of association, where they acquire the attributes of symbols and are "stored" in memory. For example, the word "bell" awakens the idea of the object it represents, but if the equivalent word of an unknown language is heard it will have significance only as a sound, without any associated idea to give it the particular meaning of bell. The process of verbal expression also consists of two parts: (*a*) impulses are first coordinated in higher association centers; (*b*) they are then transmitted to the centers that control the respiratory, laryngeal, and other muscles that play a part in the emission of speech. In written language there is a similar process, but visual receptors and centers take the part played by the auditory mechanism in spoken language, and the nerve centers and muscles of the arm and hand replace the respiratory and laryngeal nerves and muscles. The mechanism of speech is summarized in the diagram below.

Between the merely sensory and motor activities are placed the complex mechanisms of association, which endow sounds or written characters with the meaning of symbols and coordinate the complex movements needed for the emission of speech.

Sensory and motor centers occupy well-

Auditory or visual receptors, paths, and centers	→	Cortical mechanisms of association for understanding and expressing spoken or written language	→	Nerve centers, paths, and muscles of respiration, larynx, eyes, hands, etc.

Sensory or receptive mechanisms (Auditory or visual receptors, paths, and centers → Cortical mechanisms of association) | Motor or expressive mechanisms (Cortical mechanisms of association → Nerve centers, paths, and muscles)

defined areas in the cortex. Thus the auditory center is situated in the temporal lobe, the visual center in the occipital lobe, and the motor centers in the precentral gyrus. Association centers for the understanding of words or written symbols and for verbal expression are not clearly localized, but are widely spread in the cortex of the left hemisphere. They are closely correlated with intellectual capacity. Localized lesions in this widespread area produce little or no disturbance in cortical functions localized in the damaged area but may cause serious disturbances in the mechanisms of association. Lesions in the speech area, therefore, disturb the whole process of speech formation, but the disturbance is predominantly sensory or receptive when the posterior part of the area is involved and predominantly motor or expressive when the anterior part is damaged.

The speech areas are situated in the cortex of the left hemisphere in right-handed persons. They extend from the posterior part of the third, or inferior, frontal convolution through the temporosphenoidal lobe and the island of Reil to the posterior part of the parietal lobe (Brodmann's areas 4, 6*aα*, 44, 43, 41, 42, 22, 39).

Aphasia

Aphasia is a disturbance in speech due to damage in the cerebral mechanisms of verbal understanding and expression, but without mental disturbances. The lesions which cause aphasia are situated in the left cerebral hemisphere, in a large area supplied by the sylvian artery. When the lesions are situated in the posterior part of the area, defects in understanding predominate; when they are in the anterior part, defects in expression are predominant. Head maintains that all the components of speech are disturbed in aphasia, although in different cases some are more affected than others. He classified aphasia into four clinical types: verbal, nominal, syntactic, and semantic.

Verbal aphasia. Word formation is defective; there is difficulty in evoking and pronouncing words, and it is almost impossible to express ideas. Writing is difficult. This type corresponds approximately to motor aphasia (inability to speak articulately) and agraphia (inability to write).

Nominal aphasia. Words and phrases are enunciated, but there is difficulty in finding the correct words to express an idea or name an object. Reading and the mental process of arithmetic are also impaired.

Syntactical aphasia. Isolated words are sometimes formed correctly, but sentences are not properly constructed. The patient talks volubly in an unintelligible jargon without realizing that he has spoken incorrectly. Such patients are often considered insane.

Semantic aphasia. Words and short sentences are correctly understood and formed, but conversation cannot be kept up because its general significance is lost.

Déjerine's[1] classification of aphasia, derived from Charcot's theories on the subject, gives a good idea of the classical concept of disturbances in speech. Aphasia was divided into motor, sensory, and total aphasia.

Motor, or Expressive, Aphasia

Motor aphasia or Broca's aphasia. Disturbances in the formation of the spoken or written word predominate; there are few disturbances on the receptive side of the speech mechanism. Mental speech is impaired. There is no paralysis. There are lesions in the posterior part of the third or inferior frontal convolution, on the left side (Broca's center) in front of the cortical motor centers of muscles taking part in voice formation.

Pure motor aphasia. Words cannot be articulated, although mental speech is not disturbed and there is no paralysis. It is an attenuated form of Broca's aphasia.

Sensory, or Receptive, Aphasia

Sensory, or receptive, or Wernicke's aphasia. The subject does not understand the spoken word (verbal deafness) or the written word (verbal blindness), although he can hear and see. He is in the situation of a person hearing an unknown language. There is also slight impairment in the formation of words, and printed type cannot be copied into writing. Lesions of the convolutions around the posterior part of the sylvian fissure cause this type of aphasia.

In pure verbal deafness and pure verbal blindness, the patient does not understand the spoken or the written word, but there is no other defect in speech. Pure agraphia, *i.e.*, the loss of the ability to write, without paralysis or any other defect in speech, was not admitted by Déjerine.

[1] Déjerine, J., "Séméiologie des affections du système nerveux," Masson et Cie, Paris, 1914.

TOTAL APHASIA

In this type there are sensory and motor disturbances in speech. Pierre Marie[1] made a critical study of Déjerine's concepts and classification of aphasia. He concluded that only Wernicke's aphasia, due to intellectual deficiency, could be accepted as a separate entity. Pure verbal deafness and pure motor aphasia caused by lesions in the third frontal convolutions were also discarded as entities, because cases of Broca's aphasia could occur without any damage to this area, and lesions in Broca's center did not necessarily provoke aphasia. On the other hand, Marie defined a new entity, called "aphemia," due to subcortical lesions under the island of Reil. Déjerine's and Marie's classifications of aphasia are compared in Table 114.

Table 114. Déjerine's and Marie's Classifications of Aphasia

Déjerine	*Marie*
Motor aphasia	
Broca's aphasia (pure motor aphasia + verbal blindness)	Broca's aphasia (aphemia + attenuated Wernicke's aphasia)
Pure motor aphasia	
Sensory aphasia	
Wernicke's aphasia (verbal deafness + verbal blindness)	Wernicke's aphasia
Pure verbal deafness	
Pure verbal blindness	Alexia
Total aphasia	
Broca's + Wernicke's aphasia	

Aphasia is now usually considered as a severe disturbance in cortical function, which involves the whole mechanism of acquired speech, but the symptoms vary according to whether the lesions are situated mainly in the motor, auditory, or visual areas.

DEAF-MUTISM

This term is used for cases of congenital or acquired deafness in which speech has not developed. Mutism without deafness occurs very rarely. Deaf-mutes can be taught to speak, in which case they are simply deaf persons.

Historical note. Spartan law prescribed that deaf-mutes should be killed by being cast from Mount Taiget. According to Aristotle, "those born deaf then become mute," and the general idea was that these persons could not learn to speak. Following Fr. P. Ponce de Leon (1510–1584), deaf-mutes were taught to speak by several teachers in the course of the fifteenth and sixteenth centuries. Organized treatment of deaf-mutes was first started in France by Charles Michel de l'Epée (1712–1789) who taught them to express themselves by hand signs (deaf-and-dumb alphabet). Shortly afterward Samuel Heinecke (1729–1790) initiated the oral method of education in Germany, and it was soon universally adopted. Alexander Graham Bell (1847–1922), a Boston teacher of deaf persons, made many contributions to the relief of the deaf. His most outstanding achievement is, of course, the telephone. The sums he was awarded as prizes for his discoveries were dedicated to founding institutions for the benefit of the deaf and deaf-mutes.

Congenital deafness. Anomalies in development of the ear or auditory paths and centers and fetal diseases cause congenital deafness. Individuals with such deafness are mute.

Acquired deafness. Disease and trauma of the ear or auditory paths and centers occurring after birth but before the age of three years have the same effect on speech as congenital deafness. If the damage occurs at a later age, the defect in speech is not so severe, and after the age of seven or eight years there are only defects in enunciation and modulation of the voice. Deafness does not involve intellectual deficiency.

The number of the deaf-mutes is considerable, especially in certain districts (*e.g.*, areas of endemic goiter). In civilized countries there are approximately 8 deaf-mute children for every 10,000 (Hartmann). Deafness is a serious problem in public welfare. In 30 per cent of the cases there is some degree of hearing, which is of great value in learning speech, especially if it is improved by modern hearing aids.

Problems created by deafness and their treatment. The physician plays an important part in the prevention and cure of diseases causing deafness and in directing the education of the deaf. Where proper measures are taken, the number of the deaf is reduced and their disability is diminished.

Congenital deafness or deafness acquired in early childhood creates the serious problem of mutism, which isolates the patients from the community and deprives them of educational opportunities. Several methods have been invented to endow them with a conventional

[1] MARIE, P., Révision de la question de l'aphasie, *Sem. med.*, p. 241, 1906.

language. The manual method invented by de l'Epée and the oral method of Heinecke have been greatly developed and are widely used in the education of the deaf.

1. The manual method uses a special alphabet consisting of signs made by one or both hands. Words are formed, as if these signs were letters. It is therefore a kind of written language. The drawback of this type of language is that it cannot be understood by the general population but only by those who have learned it.
2. The oral method teaches the deaf-mute to form phones and words with his voice. He is taught to correlate respiratory movements, movements of the lips, and speech vibrations in the thorax and larynx with the emission of vocal sounds. He thus learns how to speak, graduating the rhythm, pitch, and volume of his voice, and how to understand the spoken word by lip-reading. Training should be commenced at an early age to avoid mental and educational backwardness; usually special schools admit children from two to six years old. Persons who have become deaf in adult life also derive benefit from this type of training.

Deafness creates not only medical and educational problems but also psychological and social ones. It undoubtedly restricts the activity of the disabled person, but it does not close many useful manual and intellectual pursuits, and has its compensation in the development of greater acuity in other senses and in freedom from noise. Edison considered his deafness an advantage because it permitted a greater degree of mental concentration. Education of deaf children and adults is of great importance if the consequences of isolation, which often lead to social maladjustments, are to be avoided.

BIBLIOGRAPHY

Berry, G., "Management of the Deaf and Hard of Hearing," Fowler Nelson Company, New York, 1939.

Culver, C., "Musical Acoustics," Blakiston, Philadelphia, 1941.

Curry, R. O. L., "The Mechanism of the Human Voice," J. & A. Churchill, London, 1939.

Gemelli, A., and G. Pastori: "L'Analisi elettroacustica del linguaggio," Università Cattolica del Sacro Cuore, Milan, 1934.

Goldstein, M., "Deaf-mutism," Fowler Nelson Company, New York, 1939.

Jobson, T., J. Nagge, V. Negus, E. Fowler, D. Guthrie, and D. Dickson, Audibility of the Radio Voice, *Proc. Roy. Soc. Med.*, **37**, 263, 1944.

Kerridge, P. M. T., Recent Advances in Knowledge Concerning Hearing and Speech, *Physiol. Rev.*, **18**, 59, 1938.

Luciani, L., "Fisiologia dell'uomo," Società Editrice Libraria, Milan, 1924.

Negus, V. E., "The Comparative Anatomy and Physiology of the Larynx," Heinemann, London, 1949.

Potter, R. K., G. A. Koop, and H. C. Green, "Visible Speech," Macmillan, New York, 1947.

Pressman, J. J., Physiology of the Vocal Cords in Phonation and Respiration, *Arch. Otolaryng.*, **35**, 355, 1942.

CHAPTER 81

The Regulation of Posture

THE DIFFERENT PARTS of the body are disposed in relation to each other, and the whole body in relation to the environment, in a manner typical for each species; *e.g.*, the upright posture in man. There are also individual differences in posture, which may be as characteristic as the facial features, height, or any other anatomic trait.

In many invertebrates, posture is passive; *i.e.*, it depends on the form of the organism and the effects of external forces, such as gravity, tropisms, and the resistance of the environment. Posture is actively adopted in vertebrates; *i.e.*, it is produced, maintained, and restored by a series of coordinated reflexes. Postural reflexes are usually unconscious, unaccompanied by pleasurable or painful sensations, and they can be maintained for a long time without fatigue. In lower animals postural reactions are not greatly disturbed by removal of the cortex; a "thalamic animal," *i.e.*, an animal in which the brain stem has been cut just below the thalamus so that the midbrain centers remain intact, maintains and can reacquire a normal posture. Certain postural reactions (*e.g.*, the so-called "placing" reactions) are integrated in the cortex, and as encephalization increases, the importance of cortical integration also increases. The extensors of the limbs and the vertebral muscles play a preponderant part in the performance of postural reflexes, but many other muscles also concur.

Stimulation of proprioceptive receptors is the main source of postural reflexes—in the first place, impulses from proprioceptive receptors in muscles. Thus if the dorsal spinal roots corresponding to a limb are cut (deafferentation), the limb loses its tonus and becomes flaccid. The labyrinth and the receptors in the ligaments of the upper cervical joints play a special part in postural reflexes. Photoreceptors and tactile receptors also may give rise to postural reactions, but they are less important and are usually supplementary to proprioceptive reflexes.

MUSCLE TONUS

Posture is built on muscle tonus, which is a muscular contraction maintained by asynchronous impulses of low frequency, discharged by the spinal motor neurons. Sherrington's myotatic (μῦς, muscle, and τατος, extended), or stretch, reflex is the primary tonic reflex

THE MYOTATIC, OR STRETCH, REFLEX

"If the tendon of a healthy muscle is drawn upon by an antagonistic muscle or by the manipulations of the investigator or by the movement of a joint in response to gravity, the muscle actively resists the extending force. A muscle which has been paralysed by section of its motor-nerve or of the ventral or dorsal roots supplying it, does not actively resist and behaves like a piece of non-contractile tissue. . . . The resistance . . . is a reflex contraction, the stretch reflex" (Sherrington). The full response can be obtained after the skin of the limb has been removed or the nerves to all the muscles except the one stretched have been cut. The reflex, therefore, arises in the receptors of the stretched muscle.

The reflex nature of the response is confirmed by the fact that it can be inhibited. Thus centripetal stimulation of a cutaneous nerve of a limb suppresses contraction provoked by stretching in the quadriceps femoris or any other extensor. Inhibition can be obtained by exerting tension on the tendon of an antagonist, *e.g.*, the

biceps femoris in the case of the quadriceps (Fig. 472). Contraction of the muscle itself also inhibits the stretch reflex (autogenous inhibition).

There are two components in the stretch reflex, a "phasic reaction" or stretch movement, in which tension increases, and a "static reaction" or stretch posture, in which tension is maintained, usually declining slightly after the peak has been reached. The phasic reaction is observed when a tendon is stretched suddenly by a tap, *e.g.*, on the patellar tendon or Achilles tendon. The stretch receptors in the tendon discharge an almost synchronous volley of centripetal impulses, which builds up a central excitatory state and excites a number of motor neurons. A brief contraction follows, *i.e.*, a tendon jerk, such as the knee jerk.

FIG. 472. Myotatic reflex. Rectus femoris is stretched 8 mm. by lowering the table. The dotted line *T* indicates table fall; *M*, reaction of uninhibited muscle; *I*, reaction of muscle inhibited by stretching the tendons of the antagonistic flexors; *P*, reaction of muscle paralyzed by nerve section (broken line). Ordinate, tension in kilograms; abscissa, time in seconds. (*Creed et al., "Reflex Activity of the Spinal Cord," Oxford, New York, 1942, p. 50.*)

The nerve impulses of the stretch reflex are transmitted along large axons at high speed, in both the afferent and efferent limbs of the reflex arc. Afferent fibers of muscles are distributed in three groups, named groups I, II, and III in order of descending caliber, with diameters about 16 μ, 8 μ, and 4 μ respectively.[1] Muscle spindles and Golgi tendon organs are innervated by group I fibers, which have the highest velocity of conduction. Stimulation of these fibers evokes monosynaptic reflex discharges with the same latency as the reflex discharges evoked by stretching the muscle. Moreover, these monosynaptic discharges are evoked only in the motor neurons corresponding to the muscle innervated by the afferent fibers which have been stimulated, *i.e.*, they have the same characteristics as those producing the stretch reflex. The myotatic reflex is therefore a monosynaptic reflex.[1]

A record of the electrical activity of the muscle shows that when it is completely relaxed there are no action currents. If the muscle is stretched slightly, action currents are recorded, beginning at a low rate of about 7 per second; the rate of firing increases as greater tension is exerted on the tendon. This effect is due to an increase in the rate of discharge in the motor units engaged, a fact that can be demonstrated by cutting down the ventral roots so that only one or a few motor units can respond. On increasing the tension, the frequency of the impulses rises from 7 to about 20. If the tension is increased even more, units hitherto at rest are "recruited" and become active, but the different motor units discharge asynchronously; therefore the apparent frequency may rise to 200 or 250 per second, although the actual frequency of discharge of any one neuron is only 30 or 40 per second. If the tension exerted on the tendon is decreased, the units are "disbanded," *i.e.*, they gradually fall out of activity and the strength of the contraction diminishes.

In spite of the low frequency of discharge, the muscle as a whole contracts smoothly. Oscillations are not apparent, owing to three factors: (*a*) the muscle fibers of a motor unit are not bunched together, but are distributed throughout the length of the muscle, so that the tension developed by the different fibers is also distributed throughout the whole muscle; (*b*) asynchronous activation of the different units contributes to damp the oscillatory effect of the subtetanic contraction of each unit; (*c*) muscle fibers taking part in sustained myotatic reflexes are mostly of the "slow" type, *i.e.*, they have a long contraction time (100 to 200 msec.); therefore complete tetanus is provoked by discharges of lower frequency than those necessary to provoke it in faster fibers.

The threshold of the stretch reflex and the strength of the contraction vary in different muscles. They are more highly developed in muscles that counteract the effect of gravity than in others, *e.g.*, the extensors of the limbs. These are

[1] REXED, B., and P. O. THERMAN, *J. Neurophysiol.*, **11**, 133, 1948; LLOYD, D. P. C., and H. T. CHANG, *J. Neurophysiol.*, **11**, 199, 1948.

[1] LLOYD, D. P. C., *J. Neurophysiol.*, **6**, 293 and 317, 1943.

"red," "slow" muscles, with abundant sarcoplasm and long contraction time.

In "fast" muscles, *e.g.*, the flexors, the phasic reaction is well in evidence, but the static reaction is not. When the tendon is stretched the muscle contracts, but soon it relaxes. Tendon jerks or "pluck" reflexes are elicited, but sustained tension (stretch posture) is not observed.

An outstanding feature of stretch reflexes is the restriction of the response to the muscle, or the part of the muscle, on which tension is exerted.

Stimulation of group I afferent fibers not only evokes discharges from the motor neurons corresponding to the muscle, or part of the muscle to which they supply afferent innervation, but also impinges directly on the immediate synergists and antagonists, producing monosynaptic facilitation in the former and inhibition in the latter. The greatest facilitatory effect is obtained when the conditioning and test volleys coincide in time; facilitation decays exponentially, falling to half in about 4 msec. The direct inhibitory effect on the antagonists, after a brief rising phase, also decays exponentially.[1] Lloyd refers to this group of muscles bound to each other into mutual dependence by monosynaptic interconnections, together with the monosynaptic links that bind them, as a "myotatic unit."[2]

Stimulation of fibers in groups II and III, which have a higher threshold and a lower velocity of conduction than group I fibers, evokes flexor responses; also facilitation or inhibition of monosynaptic reflexes by intercalated interneurons.

Lengthening and shortening reactions. Two aspects of the stretch reflex, *i.e.*, the lengthening and shortening reactions, are prominent in the decerebrate preparation. If an attempt is made to bend the knee of an animal in decerebrate rigidity, the extensor muscles resist up to a point, after which they suddenly give way and the limb can be flexed; there is a "clasp-knife" effect. This is caused by inhibition of the stretch reflex (autogenic inhibition), due to excessive tension, and is known as the lengthening reaction. If on the contrary the flexed limb of a decerebrate preparation is extended, the extensors relax for a moment and then shorten so as to hold the limb in the new position. This is the shortening reaction. Both these reactions endow the decerebrate animal with the capacity to hold a given position.

The stretch reflex is fundamentally a spinal reaction. It is observed in spinal animals, but it is modified by facilitatory and inhibitory impulses from higher centers, which are of great importance in the regulation of tonus and postural reactions. The spinal animal will be considered first, and then other preparations such as the decerebrate animal, in which the midbrain centers also remain intact.

The Spinal Animal

Spinal shock. Immediately after a transverse section of the spinal cord has been performed, the following effects are observed in the segments below the level of the section: (*a*) complete anesthesia;[1] (*b*) paralysis; (*c*) flaccidity of the muscles (loss of tonus); (*d*) suppression of somatic and visceral reflexes. Paralysis and the loss of sensation are permanent disabilities due to section of the ascending (spinothalamic) and descending (pyramidal) tracts, the fibers of which do not regenerate.

Depression of reflexes is the outstanding sign of spinal shock, but it is not permanent. It is more marked and lasts longer as encephalization increases. In amphibians it lasts a very short time, and a few minutes after spinal-cord section, reflex responses of normal strength can be elicited. In the cat the depression lasts longer, and in the dog it is even more prolonged; in primates, shock may last for several days, which lengthen into weeks in the anthropoids and man.

The first reflex to reappear after transection of the spinal cord is the flexion reflex. At first there is only a limited response to strong stimulation of the distal end of the limb. Later it increases in strength and amplitude, and eventually hyporeflexia is replaced by hyperreflexia. The knee jerk is the first extensor reflex to reappear; in the cat sometimes it is not completely suppressed, but only depressed. The phasic reaction of myotatic reflexes is recovered fairly soon, but the static reaction remains depressed for weeks, even in the cat and dog. Flaccidity, typical of spinal shock, is due to this depression of the

[1] *Ibid.*, **9**, 421, 1946.

[2] Lloyd, D. P. C., *Research Publ., A. Nerv. & Ment. Dis.*, **30**, 48, 1950.

[1] Some patients complain of pain (poorly localized) below the level of the injury. This pain is probably due to pain afferent fibers in the sympathetic which enter the cord at a higher level.

stretch reflex. The crossed-extension reflex reappears when there is already a well-developed flexion reflex. Cutaneous reflexes, such as the scratch reflex, which involve several spinal segments, are recovered later. In this sequence the early recovery of defense reflexes in response to nocuous stimuli is conspicuous. Reflexes arising in proprioceptors, with the exception of tendon jerks, reappear much later. The predominance of flexion over extension is demonstrated very vividly by transection of the spinal cord in a decerebrate animal with fully developed rigidity due to hyperactivity of the stretch reflexes. Rigidity is rapidly replaced by flaccidity and the threshold of extensor reflexes rises as the threshold for flexor reflexes diminishes.

Visceral reflexes are also depressed in spinal shock. The blood pressure falls owing to loss of vasomotor tonus, and pressor reflexes cannot be elicited. After a few days (in the dog) the blood pressure may be higher, but pressor reflexes are not recovered even after a long period of time. Sweat secretion is abolished, and the skin is dry; owing to vasodilatation, it is pink and warm.

Reflex evacuation of the bladder and rectum are not performed. Immediately after transection, the sphincters of the bladder are strongly contracted and there is retention of urine, which has to be evacuated by catheterization. After about three weeks in man, bladder reflexes are partially restored, but the sphincters tend to become incompetent. The bladder, however, is not evacuated completely. Microorganisms grow in the residual urine, which may cause inflammation of the bladder (cystitis). Reflex evacuation of the rectum is recovered after 3 or 4 weeks.

Other visceral functions are also recovered, including those of sex. Ovulation, impregnation, pregnancy, and parturition have been observed in bitches after complete spinal transection at the level of the lower cervical or upper thoracic segments.

Spinal shock is particularly severe in man.[1] All reflexes are completely suppressed for several weeks in most cases, but in some individuals reflexes from the perineum and genital area are never lost, and in others superficial reflexes reappear within 24 hr. after the cord has been cut.[2] After 2 or 3 weeks the flexion reflex reappears in response to strong stimulation of the plantar surface. At first there is only slight withdrawal, which gradually becomes more vigorous. A peculiar aspect of this response is the extension of the toes, especially of the big toe; this is known as "Babinski's sign" and is always accompanied by contraction of the hamstrings.[1] Babinski's sign is evidence of pyramidal-tract injury and is observed when this tract is damaged, even if the rest of the nervous system is uninjured. The knee jerk reappears later, and the crossed-extension reflex is not restored until 6 or 7 weeks have passed.

[1] Head, H., and G. Riddoch, *Brain*, **40**, 188, 1917.

[2] Kuhn, R. A., *J. Nerv. & Ment. Dis.*, **113**, 301, 1951.

The flexion reflex gradually becomes exaggerated, and after several months cutaneous stimuli provoke a widespread response, with complete withdrawal of the limb, profuse sweating, and evacuation of the bladder and rectum. This is known as the "mass reflex." Many months after spinal-cord transection, a slight degree of postural extension may be observed, but spasticity occurs only in cases of incomplete transection. Reflex rigidity can be provoked several months after the cord has been cut by pressure against the popliteal region. All the muscles innervated by the distal segment of the cord contract and form a rigid framework, so that the patient can stand up without support. Later the muscles relax simultaneously, and the patient collapses if not supported.

Section of the cord above the emergence of the phrenic nerves causes death by asphyxia because respiration ceases. If the section falls below the fourth cervical segment, there is quadriplegia (paralysis of the four limbs) but the respiratory movements of the diaphragm are not disturbed. The effects of shock in the lower limbs are less marked in cases of transection at a high level than when the injury involves the lower segments of the cord; frequently there is priapism (persistent erection of the penis).

The nervous mechanism of spinal shock. Spinal shock is due to the sudden suppression of impulses descending from the upper centers to the motor neurons in the spinal cord. It is not due to trauma or irritation at the level of injury, because its severity is the same whether a clean section is made or several segments are crushed. Moreover if a second section is made below a previous transection, after the animal has re-

[1] Fulton, J. F., and A. D. Keller, "The Sign of Babinski. A Study of the Evolution of Cortical Dominance," Charles C Thomas, Springfield, Ill., 1932.

covered from the effects of the first operation, spinal shock does not occur again.

In the cat, the dorsal two-thirds of the cord can be cut without provoking shock, but section of the ventral columns is followed by shock. Impulses that protect the animal from shock descend in the vestibulospinal tract and in fibers of the reticulospinal tract.[1] Shock is not provoked if the section is made above the level of the vestibular nuclei; on the other hand, destruction of these nuclei produces shock.

In primates, corticospinal impulses are of greater, and vestibulospinal impulses of lesser, importance. Only slight reflex depression follows destruction of the vestibular nuclei, while destruction of the motor cortex causes marked signs of shock.

In lower animals (*e.g.*, amphibians) spinal reflexes are carried out with little or no interference from the higher centers, but as these centers become more developed, their importance in all nervous activity increases. In primates encephalization is at its highest, and even spinal activity cannot be normally carried out without the involvement of supraspinal centers. Impulses from these centers either facilitate or inhibit spinal mechanisms, which can be depressed owing to the absence of facilitatory impulses, or enhanced owing to release from inhibition.

The spinal animal at first shows the effects of the absence of facilitating impulses descending

FIG. 473. Brown-Séquard syndrome. Effects following section of left half of the cord.

[1] FULTON, J. F., E. G. T. LIDDELL, and D. McK. RIOCH, *Brain*, *53*, 311 and 327, 1930.

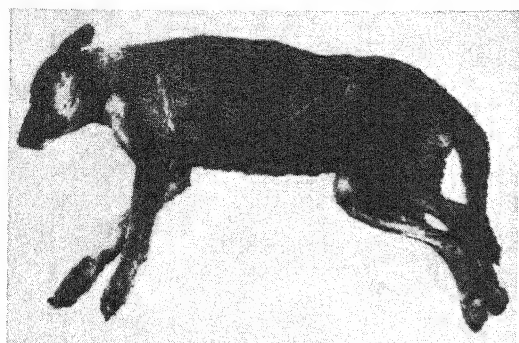
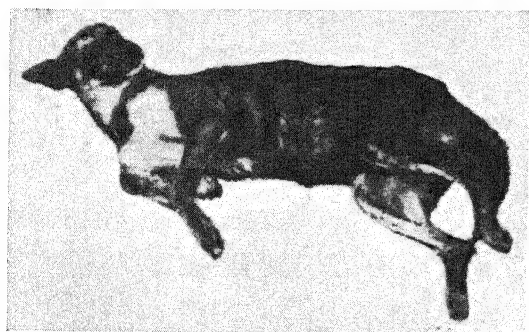
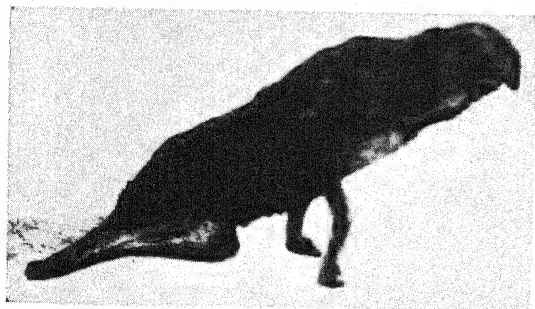
from the higher centers, and spinal reflexes are depressed. Later a more primitive, purely spinal, equilibrium is established; spinal mechanism alone can integrate simple reactions such as defensive and visceral reflexes. Later still the phenomena of release, *i.e.*, absence of inhibition, become prominent and the mass reflex is observed because the central excitatory state spreads, without order or inhibitory control, to a large number of spinal reflex centers.

a

b

c

FIG. 474. Complete section of spinal cord at upper lumbar segments. *a*, the animal lying quietly; note extension of forelimbs; *b*, the animal's attention is awakened; note flexion of forelimbs; *c*, the animal walks dragging its hindquarters and leaving a trail of urine because of incontinence of the bladder.

Incomplete transection of the spinal cord. Brown-Séquard syndrome. In 1855 the French-American scientist Brown-Séquard described the effects of lateral semisection of the spinal cord (Fig. 473). The following effects are observed in the segments below, and on the same side as, the lesion: (*a*) paralysis, *i.e.*, loss of voluntary movements, due to section of the pyramidal tract; (*b*) loss of proprioceptive sensibility due to section of the dorsal column (cuneate and gracile fasciculi); (*c*) transitory cutaneous hyperesthesia; (*d*) at first, depression of spinal reflexes owing to the absence of facilitatory impulses mediated by the vestibulospinal and pyramidal tracts, and later release phenomena (hyperreflexia); (*e*) at the level of the lesion, if several segments have been involved, an area of complete anesthesia owing to destruction of the dorsal roots, and above it an area of hyperesthesia corresponding to the incompletely denervated dermatome; (*f*) transitory increase in skin temperature due to cutaneous vasodilatation.

On the side opposite the lesion there is cutaneous anesthesia due to section of the spinothalamic tract. Temperature sensations and pain are more severely involved, while tactile sensations are not disturbed, except in a few segments immediately below the lesion. Proprioceptive sensibility and voluntary and reflex movements are normal.

Hemisection of the dorsal quadrants of the spinal cord causes the loss of proprioceptive sensations (section of the dorsal columns) and paralysis (section of the pyramidal tract), usually accompanied by spasticity. The limbs are extended, not flexed as occurs after the signs of shock have receded in cases of complete transection. Extensor spasms can be provoked only if the vestibulospinal tract and the corticospinal paths have not been completely severed.

Lesions that involve only one ventral quadrant produce cutaneous anesthesia (mainly the suppression of thermal sensations and pain) on the side opposite to the injury. This is due to section of lateral spinothalamic pathways. There is also a tendency to paresia and flaccidity due to section of the vestibulospinal and part of the corticospinal tracts.

Schiff-Sherrington phenomenon. A dog or a cat, after complete transection of the cord at

the level of the thoracic segments, shows signs of disturbed function not only below the lesion, but also in the segments above it. There is usually an increase in extensor tonus in the forelimbs; the animals lie with their hind limbs flaccid and the forelimbs spastically extended (Fig. 474). This phenomenon is partly due to release from inhibitory impulses arising in the hind limbs and integrated in the cerebellum. The following facts are evidence of this assertion:[1] (*a*) hind-limb deafferentation in decerebrate animals converts into rigidity the flaccidity of a deafferented forelimb; (*b*) transection of the spinal cord at the thoracic level, or extirpation of the anterior lobe of the cerebellum, has the same effect; (*c*) ipsilateral labyrinthectomy immediately suppresses this rigidity. Tonic impulses from the labyrinth to the forelimb are, therefore, inhibited by impulses from the cerebellum which arise in the receptors of the hind limb. The Schiff-Sherrington phenomenon is not, however, exclusively due to release from cerebellar inhibition, because rigidity in the deafferented forelimb has been observed after extirpation of the cerebellum. Moreover, if the cord is cut at the level of the thoracic segment and later a second transection is made between the lumbar and sacra segments, the knee jerk and postural tonus of the quadriceps (integrated at the fourth and fifth lumbar segments) are increased. There is therefore a spinal component in this phenomenon. Impulses from the hind-limb receptors are not the only source of inhibitory impulses, because extension of the forelimbs is also observed if the cord is cut after lumbosacral deafferentation.[2]

Decerebrate Rigidity

If the brain stem of a cat or a dog is cut between the superior and inferior colliculi, marked rigidity develops (decerebrate rigidity). In whatever position the animal is placed, its four limbs are kept rigidly extended. The head, the spinal column (opisthotonus), and the tail are also extended. The jaws are firmly clamped. The animal cannot stand up, since it has lost the righting reflexes, but if it is placed on its feet, it stands on its toes.

Deafferentation of a muscle (section of the corresponding dorsal roots) suppresses rigidity in that muscle, but not in others. Rigidity therefore arises in the muscle receptors (muscle spindles) and is simply an exaggeration of the myotatic reflexes, especially developed in those muscles which counteract the effects of gravity. The deafferented muscles can, however, respond reflexly to impulses arising in other receptors, *e.g.*, they can take part in the crossed-extension reflex and in labyrinthine reactions.

Rigidity can be suppressed by inhibition. Thus stimulation of the motor nerve of an antagonistic muscle, *e.g.*, the biceps femoris, causes the quadriceps to become flaccid. The same effect is observed after stimulation of a cutaneous nerve, which provokes a flexion reflex.

Rigidity is more marked in muscles that counteract the effects of gravity, but it is also evident in the flexors, which respond to a pull on the tendon with clonic (repeated) contractions, not simply by a "pluck" reflex. The threshold of flexor reflexes is, however, high, while that of extensor reflexes is low, in the decerebrate preparation—a condition opposite to that observed in the spinal animal.[1]

The nervous mechanism of decerebrate rigidity. Rigidity is due to exaggerated stretch reflexes owing to suppression of inhibitory impulses and predominance of facilitatory impulses descending from centers above the spinal level. Thus transection of the ascending tracts in the dorsal columns of the spinal cord does not modify rigidity. Transection of the anteroventral column suppresses it on the same side in the segments below the section. Magnus endeavored to localize the centers responsible for rigidity in the cat by transecting the brain stem at different levels. Rigidity developed when the section was made between the anterior and posterior colliculi; sections at higher levels did not provoke it. Sections at lower levels did not modify rigidity until the vestibular nuclei were excluded, in which case it was abolished, the preparation becoming flaccid.

Transection between the colliculi excludes the red nucleus, and rigidity may be observed when only the parvicellular portion of the red nucleus has been excluded. Complete destruction of the red nuclei or their descending projections by

[1] Stella, G., *Boll. Soc. ital. biol. sper.*, **22**, 78 and 81, 1946.

[2] Ruch, T. C., *Am. J. Physiol.*, **114**, 457, 1935.

[1] The postural significance of decerebrate rigidity is evident in the three-toed sloth. This animal normally counteracts the effects of gravity with its flexor muscles. Decerebration causes rigidity to develop in these muscles and exaggerated flexion is observed (Richter, C. P., and L. H. Bartemeier, *Brain*, **49**, 207, 1926).

means of restricted lesions does not, however, provoke rigidity in the cat. The animals show signs of disturbance in cerebellar function (see Chap. 83), and a mild degree of extensor hypertonus in certain conditions, but "decerebrate rigidity" is not provoked by this operation.[1] The red nucleus alone, therefore, cannot be responsible for the inhibitory impulses that prevent the development of rigidity.

Magoun[2] and his associates have demonstrated the importance of inhibitory influences from the bulbar reticular formation. In an anesthetized animal with a background of motor activity, stimulation of the reticular formation at the bulbar level, especially in the ventromedial area, produces complete cessation of movements. Two-neuron stretch reflexes, multineuron flexor movements, and movements originated in the cortex are inhibited. This suppressor effect has a low threshold, and the impulses are conducted at a high velocity. The descending path is found in the anterolateral column of the spinal cord, and it ends in the lateral portion of the intermediate interneuron pool.[3] This path is for the most part ipsilateral, but there is some crossing over at the spinal level. Facilitation can also be obtained by stimulation of the reticular formation, at the bulbar level laterally to the inhibitory areas,[4] but more especially above the bulbar level up to the diencephalon (mid-line and intralaminar nuclei of the thalamus, subthalamus, and hypothalamus). Facilitation affects all types of movement; it has a low threshold, and the impulses are conducted in fast fibers. The descending path has several relays; in the spinal cord it overlaps the inhibitory pathway. Direct endings on the spinal motor neurons give only minor facilitation, but the descending fibers give out many large collaterals to interneurons with short fibers, which activate many motor neurons and moreover recruit other interneurons, thus facilitating and synchronizing discharges from a large number of motor neurons.

Extirpation of the frontal cortex is followed by rigidity, especially marked in primates. Re-

[1] INGRAM, W. R., and S. W. RANSON, *Am. J. Physiol.*, **102**, 466, 1932.

[2] MAGOUN, H. W., *Physiol. Rev.*, **30**, 459, 1950.

[3] MAGOUN, H. W., and R. RHINES, *J. Neurophysiol.*, **9**, 219, 1946; SPRAGUE, J. M., *et al.*, *J. Neurophysiol.*, **11**, 501, 1948.

[4] RHINES R., and H. W. MAGOUN, *J. Neurophysiol.*, **9**, 219, 1946.

moval of the premotor cortex alone has the same effect. Inhibitory impulses travel in fibers of the pyramidal tract, because stimulation of this tract suppresses rigidity provoked by transection of the midbrain below Forel's decussation; stimulation of the contralateral frontal cortex in these circumstances inhibits rigidity.[1] Inhibitory fibers have been traced descending from area 4*s* (see Chap. 82) in the internal capsule, down to the lower level of the pons where partial decussation takes place; these fibers end in the bulboreticular formation, where impulses that inhibit not only the myotatic reflex but also movements of cortical origin[2] are relayed to the motor centers of the spinal cord. Similar inhibitory impulses originated in areas 19*s* and 24*s* travel to the suppressor areas of the bulbar reticular formation.[3] Facilitatory impulses from area 6 in the frontal lobe are relayed in the midbrain tegmentum.[4] Depression of the electrical activity of the brain by intravenous injection of sodium cyanide is followed by a transient period of decerebrate rigidity, during which the electrical activity of the suppressor reticular formation is reduced, while the facilitatory reticular formation remains active. This effect is due to failure of inhibitory impulses from the cortex and persistence of facilitatory impulses from the tegmentum of the pons. If the pontile tegmentum has been previously damaged, cyanide injection does not provoke rigidity.[5]

Extirpation of the cerebellum or section of the superior cerebellar peduncles increases rigidity. An inhibitory pathway has been traced from the Purkinje cells in the anterior lobe and paramedium lobule of the cerebellar cortex to the fastigial nucleus, and thence to the bulboreticular formation.[6] Inhibitory impulses are not discharged from the fastigial nucleus after the cerebellar cortex has been removed. Spasticity can be produced by suppression of a single one of the inhibitory cortical areas, but the effects

[1] The pyramidal tract decussates at a lower level and is not cut by the section.

[2] McCULLOCH, W. S., C. GRAF, and H. W. MAGOUN, *J. Neurophysiol.*, **9**, 127, 1946.

[3] WARD, A. A., *J. Neurophysiol.*, **11**, 13, 1948; McCULLOCH, W. S., and E. HENNEMAN, *Federation Proc.*, **7**, 79, 1948.

[4] PETERSON, E. W., and D. S. BICKERS, *Tr. Am. Neurol. A.*, 168, 1949.

[5] WARD, A. A., *J. Neurophysiol.*, **10**, 89, 1947.

[6] SNIDER, R. S., H. W. MAGOUN, and W. S. McCULLOCH, *Federation Proc.*, **6**, 207, 1947.

are transient. More pronounced and permanent effects are obtained by combined removal of cerebral and cerebellar inhibitory areas.[1]

In primates, including man, two types of rigidity have been described: (*a*) transection at a low mesencephalic level causes rigidity with the four limbs and the head extended (decerebrate rigidity); (*b*) transection at a higher level, or removal of the frontal lobe, causes rigidity with the upper limbs flexed and the lower limbs extended (decorticate rigidity) (Fig. 475).

Fig. 475. Magnus and de Kleijn reflexes in subject with "high" decerebration. On rotation of the head, the limbs on the cranial side are flexed, and those on the mandibular or facial side are extended. With the head in the "normal" position, *i.e.*, facing forward, the upper limbs are flexed and the lower limbs are extended.

Decerebrate rigidity is a classic example of a "release phenomenon." During the last century Jackson postulated that the "positive" symptoms of hemiplegia due to hemorrhage in the internal capsule (*i.e.*, hyperreflexia and spasticdity) were not caused by irritation of nerve paths, because they were permanent disabilities, but were caused by "release" of the lower centers from inhibitory impulses normally descending from the higher centers. This inhibitory effect is mediated mainly by the pontobulbar reticular formation. Spasticity caused by the destruction of the corticobulboreticular suppressor system is due to the unbalanced effect of the facilitatory system. The midbrain tegmentum is the main facilitatory center, because lesions above this level do not modify spasticity, but injury of the tegmentum reduces it considerably. Destruction of the vestibular nuclei does not modify spasticity when the midbrain reticular formation is active, but abolishes it when combined with removal of the latter.[1]

The diencephalic-midbrain facilitatory and suppressor systems also regulate the level of general cortical activity (see Chaps. 87 and 88).

Regulation of Muscle Tonus

Muscle tonus is fundamentally a reflex arising in the muscle itself, *i.e.*, the stretch reflex, which is more prominent in muscles counteracting the effect of gravity. The stretch reflex is facilitated by impulses descending from the motor cortex (area 4) the tegmentum and the vestibular nuclei. It is inhibited by impulses descending from the premotor cortex (4*s*) and the cerebellum, mainly through the reticular formation to the spinal motor centers. Suppression of facilitatory impulses causes depression of tonus for a certain time (*e.g.*, in spinal shock). Later a new equilibrium, in which flexor responses are predominant, is

[1] Lindsley, D. B., L. H. Schreiner, and H. W. Magoun, *J. Neurophysiol.*, **12**, 197, 1949.

[1] Schreiner, L. H., D. B. Lindsley, and H. W. Magoun, *J. Neurophysiol.*, **12**, 207, 1949.

established in the spinal cord. Suppression of inhibitory impulses is followed by an increase in tonus if the facilitatory impulses are active; there is rigidity, especially in muscles that counteract the effects of gravity, and extensor reflexes predominate.

The different centers that regulate tonus act not individually but coordinately, and their impulses are summated when they have similar effects. Spinal centers, vestibular nuclei, the tegmentum, and the motor cortex have a positive influence on tonus. Mesencephalic centers, the cerebellum, and the premotor cortex have a negative (inhibitory) influence. Muscle tonus is regulated by the interplay of the two mechanisms.

POSTURAL REFLEXES

Magnus classified postural reflexes into (*a*) static reactions and (*b*) statokinetic reactions. Static reactions come into play when the body is resting; they are responsible for the establishment and maintenance of posture. They may be limited to a single limb (local static reactions) or may involve the whole segment (segmental static reactions). More complex responses are carried out by more than one spinal segment (intersegmental reactions) or they may also involve supraspinal centers (general static reactions). Statokinetic reactions are observed when the body is moving, *e.g.*, walking or running.

Local static reactions. "A movable limb is at times used as an instrument for very different purposes (such as scraping, scratching, fighting, etc.) and moves freely in all joints, whereas at other times it is transformed into a stiff and strong pillar, which gives the impression of being one solid column, able to carry the weight of the body."[1] The coordinated reflex responses that fix the joints into a column constitute Magnus's "positive supporting reaction." The column is broken down, and the limb set free so that it can be moved, by another series of reflexes known as the "negative supporting reaction." The stretch reflex plays a fundamental, but not exclusive, part in the formation of the column; there is also synergical contraction of extensors and flexors, abductors and adductors, which in this reaction do not act as antagonists as they do in phasic reactions.

The positive supporting reaction is initiated in two kinds of receptors; (*a*) proprioceptors in the small muscles of the foot, *i.e.*, stretch receptors in the interosseus muscles stimulated by separation of the toes, and in plantar flexors stretched by dorsal flexion of the foot; (*b*) exteroceptors, *i.e.*, tactile receptors in the skin of the plantar surface. In the decerebellate dog, Magnus saw that a light tactile stimulus on the plantar surface of the toe pads provoked extension, so that the foot seemed to be attracted by the observer's stimulating finger as by a magnet ("magnet reaction"). This reaction is reduced to the extensor thrust provoked by pressure exerted on the plantar surface, in the spinal preparation. When the animal stands on its feet, both types of receptor are stimulated and the effects are summated. An increase in the load on the supporting column increases the strength of stimulation and provokes a more vigorous reflex response, which reinforces the column.

The negative supporting reaction that releases the limb is initiated by (*a*) suppression of stimuli for the positive supporting reaction, *e.g.*, taking the load off the limb; (*b*) inhibition of the positive reaction by plantar flexion or by the application of a stimulus that provokes a preponderant phasic reflex, *e.g.*, defensive withdrawal of the limb.

Segmental static reactions. The crossed-extension reflex is one example of a postural reaction that involves both sides of a segment. Another is reflex extension, produced by stretching the adductors of the opposite limb. This occurs when the weight of the body is supported on one foot and the subject leans to the opposite side; the adductors are stretched and the opposite limb is extended, thus preventing the body from falling.

Intersegmental static reactions. The effects on the limbs of flexion or extension of the trunk are good examples of postural reactions carried out by several spinal segments. Flexion of the trunk increases extensor tonus in the hind limb and diminishes it in the forelimb; extension of the trunk has the opposite effect.

General static reactions. The position of the head is a predominant factor in reflex stance. Reflexes are initiated in (*a*) the proprioceptors in the neck; (*b*) the labyrinth. Both types of receptors are stimulated by changes in the position of the head, and they produce effects that reinforce each other. Labyrinthine reactions can be suppressed by destroying the labyrinth, or the neck reflexes can be suppressed by immobi-

[1] Magnus, R., *Lancet*, **2**, 531 and 585, 1926.

lizing the neck in a plaster cast. In order to avoid the effects of the righting reactions, decerebrate animals are used in the analysis of these phenomena.

Tonic neck reflexes. These reflexes can be examined in the decerebrate preparation after bilateral labyrinthectomy. The following effects are observed:

1. Rotation of the head produces an increase in extensor tonus in the forelimb and hind limb on the side toward which the face is turned and a decrease in extensor tonus (flexion) on the opposite side.
2. Inclination of the head toward one side increases extensor tonus on that side and decreases it on the opposite side.
3. Extension of the head by bending the neck backward increases extensor tonus in the forelimb and diminishes it in the hind limb.
4. Flexion of the head increases extensor tonus in the hind limb and diminishes it in the forelimb.
5. Pressure exerted on the spinous process of the last cervical vertebra diminishes extensor tonus (*i.e.*, produces flexion) in the four limbs.

Section of the three upper cervical dorsal roots suppresses these reflexes. Tonic neck reflexes do not arise in the neck muscles, because they can be evoked after bilateral section of all the muscular and cutaneous branches of the first three cervical nerves and denervation of all the muscles in the neck. They arise in the ligaments of the upper cervical joints, especially the occipitoatlantal, and atlantoaxial joints.[1]

Tonic labyrinthine reflexes. These reflexes can be observed without the complicating effects of neck and other reflexes in decerebrate animals with the neck fixed in a plaster cast, or after the three upper cervical dorsal roots have been cut. The following are the main reactions:

1. Lateral inclination of the head increases extensor tonus in both limbs on the side toward which the head is bent and diminishes it on the opposite side.
2. When the animal is placed on its back, the extensor tonus of the four limbs increases; maximal effects are observed when the mouth cleft is inclined upward at 45° to the horizontal.
3. When the animal is placed on its belly, extensor tonus decreases in the four limbs; maximal effects are observed when the mouth cleft is inclined downward 45° to the horizontal, *i.e.*, the animal has rotated on a transverse axis 180° from the position it occupied when extensor tonus was at a maximum.
4. Rotation on a vertical (longitudinal) axis does not provoke static labyrinthine reactions, because the receptors (otolith organs) are not stimulated by changes in position that take place in a horizontal plane.

[1] McCouch, G. P., *et al.*, *J. Neurophysiol.*, **14**, 191, 1951.

Labyrinthine reflexes have a direct effect on muscle tonus and an indirect effect produced by stimulation of the receptors in the neck when the head changes its position. In the intact animal, the effects of labyrinth and neck reflexes are summated when the head is inclined sideways. When the head is flexed or extended, the effects are summated in the forelimbs, but are of opposite sign in the hind limbs. The total effect will therefore be greater on the forelimbs than on the hind limbs.

Static labyrinthine reflexes are due to the position and not to movements of the head; bilateral destruction of the labyrinth suppresses them.

Cases of brain tumors have been observed in man in which the connections of the cortex with the lower centers are interrupted. Such patients are the equivalent of decorticate or thalamic animals. Tonic neck and labyrinth reflexes are prominent features in these subjects (Fig. 475).

Righting reactions. Normal animals acquire and maintain posture by a series of reactions initiated in several types of receptor. These reflexes can best be analyzed in the decorticate or thalamic animal, *i.e.*, after all cortical impulses have been suppressed. The different receptors are then inactivated except for the one that is under observation. The animal is in the "zero condition" when all stimuli capable of provoking postural reflexes have been eliminated. This condition is achieved when the animal has its neck fixed in a plaster cast or the upper cervical roots cut, has both labyrinths destroyed, and is held in the air taking care to exert pressure symmetrically on the body. If the cortex is intact, cats, dogs, and primates must be blindfolded, because in these animals visual stimuli are im-

portant factors in righting reactions. The following are the righting reflexes:

Labyrinthine righting reflexes. If an animal is held in the air the head tends to regain the normal position. After bilateral labyrinthectomy or section of the vestibular branch of the eighth nerve, the head hangs loosely under the action of gravity.

Body righting reflexes. If pressure is exerted on one side of the body of a labyrinthectomized animal, the head tends to assume the normal position. If the pressure is exerted symmetrically on both sides, the head hangs loosely.

Neck righting reflexes. When the head is turned to the normal position, the proprioceptors in the neck are stimulated and reflexes righting the thorax and pelvis are initiated.

Reflexes from the trunk acting on the body. If an animal in which labyrinthine and neck reflexes have been suppressed is placed lying on one side, the head and body are righted. The body is righted even if the head is fixed and prevented from returning to the normal position. This is due to pressure exerted on the undersurface of the body, as can be demonstrated by placing a board on the upper side of the animal and exerting a pressure equivalent to its weight; in this case the animal does not right itself.

Visual righting reflexes. In primates, and in cats and dogs, visual stimuli are of considerable importance in righting reactions, and the normal position can be restored in these animals even if all other stimuli for postural reflexes are suppressed. These reflexes are not observed after the occipital cortex has been removed, therefore they are not seen in thalamic animals.

Placing reactions. The feet are placed in a position suitable for normal standing by means of the so-called "placing reactions,"[1] which are initiated by visual, cutaneous, and proprioceptive stimuli. For example, when an animal is moved downward toward a supporting surface the limbs are stretched so as to support the body. This reaction is due mainly to visual stimuli (visual placing), but a blindfolded animal will also present it if it is moved downward at a speed sufficient to stimulate the labyrinth (vestibular placing). If the visual and vestibular placing reactions are suppressed, the animal "places" its foot as soon as the slightest contact is established between the limb and the edge of the table (contact placing). Extirpation of the cortex suppresses placing reactions on the opposite side of the body. In the cat and monkey the cortical centers are found in the motor cortex. In the monkey placing reactions persist after extensive cortical extirpation if the motor area (4) remains intact; destruction of this area abolishes placing reactions (Bard).

Hopping reactions. If an animal having its body supported by a single limb is displaced in a horizontal plane, the leg hops in the direction of the displacement so that it is kept below the body and can continue to support the animal. The reaction is initiated by stretching of the adductor muscles, and it is temporarily depressed by removal of the cortex.

Compensatory eye movements. When the position of the head is altered, the eyes move so as to keep the eye fields approximately unchanged. The reactions consist of two phases: (*a*) the eyes are first placed in the new position by contraction and relaxation of the corresponding muscles; (*b*) the eyes are then tonically fixed in the final position. Stimulation of the semicircular canals initiates the statokinetic phase, and stimulation of the utricle is responsible for the tonic or static phase. Impulses from proprioceptors in the muscles of the neck have the same effect as, and summate with, the labyrinthine reflexes.

The following are the principal reactions: (*a*) flexion of the head provokes contraction of the superior rectus and inhibition of the inferior rectus; (*b*) extension of the head has the opposite effects; (*c*) rotation of the head provokes contraction of the medial rectus of the eye on the side toward which the head is turned and of the lateral rectus of the other eye; the corresponding lateral and medial rectus muscles are relaxed. The utricle does not play any part in the latter reaction, because it is not stimulated by changes in the position of the head which take place in the horizontal plane.

Statokinetic reactions. These reactions are produced by movement, *i.e.*, positive or negative acceleration (increase or decrease in speed). They are observed when an animal moves in a horizontal plane or rotates around an axis. Thus when an animal is lifted up from a supporting surface it flexes its limbs and flexes the head; when it is moved downward, or dropped, the head and limbs are extended. When it is rotated, a series of compensatory movements take place in the eyes, head, and trunk which are

[1] Bard, P., *Harvey Lect.*, 33, 143, 1937–1938.

known as nystagmus. When the head rotates, the eyes move slowly in the direction opposite to that of rotation, tending to keep the eye fields unchanged, then by a quick movement the eyes are brought back to the normal position, *i.e.*, looking straight forward.[1]

Nystagmus may be horizontal, vertical, or rotatory according to whether its movements take place in the horizontal, sagittal, or frontal plane.

When the body rotates, the head turns slowly in a direction opposite to that of rotation and is then brought rapidly to the normal position; thus a "head nystagmus" results, with the rapid component in the same direction as that of rotation. The trunk also performs compensatory movements that tend to keep it in the normal position.

If rotation is kept up for a sufficient time, nystagmus ceases, because it is due to acceleration, not to movement in itself. When rotation ceases, nystagmus is again observed (postrotatory nystagmus) but in the opposite direction to that which occurred during rotation. The subject experiences a sensation of falling, and unnecessary compensatory movements are performed, which cause him to fall (see page 1032).

Changes in posture stimulate certain exteroceptive and proprioceptive receptors, and thus initiate compensatory static and statokinetic reactions. At the same time stimulation of other receptors ceases, and the postural reflexes that maintained the previously existing pose are suppressed. There are therefore positive (excitatory) and negative (inhibitory) postural reactions. The result of all these reactions is a posture that keeps the body in equilibrium under many different circumstances.

Centers that coordinate postural reactions. The "thalamic" animal can perform nearly all postural reflexes. It can therefore not only maintain a posture, but also right itself if placed in an abnormal position. (Lesions that reproduce in man the condition of the thalamic animal do, however, disturb postural reflexes, and the erect posture cannot be maintained.) Transection of the brain above the emergence of the third nerve does not disturb postural reflexes; transection below this point abolishes them. The cerebellum and the dorsal part of the midbrain can be removed without causing more than transitory disturbances in posture. Lesions in the ventral part of the midbrain cause severe alterations in postural reactions. The substantia nigra and the red nuclei can be destroyed without causing serious disturbances in postural reactions; the animal remains capable of righting itself and walking, even when it is blindfolded. Centers in the reticular formation have been shown to be of importance in the integration of postural reactions.

The cortex also plays a part in the integration of posture; thus optic impulses that give rise to postural reflexes are integrated in the occipital cortex, and the placing and hopping reactions are suppressed by extirpation of the motor cortex.[1] Ablation of the parietal cortex causes only a transitory depression of these reactions. Centripetal stimulation of the dorsal roots provokes electrical activity in the postcentral convolution which is not of tactile origin.[2] Stimulation of proprioceptors in cats and monkeys by means of passive movements or stimulation of a motor nerve provoked electrical activity in the cortex. In the cat the sensory motor area on the opposite side (sometimes on both sides) was activated; the temporal, parietal, and occipital lobes showed no changes. In the monkey areas 4 and 6 became active.[3]

FUNCTIONS OF THE LABYRINTH

The labyrinth consists of two parts: (*a*) the utricle and saccule; (*b*) the semicircular canals.

The maculae in the saccule and utricle are receptor organs innervated by the vestibular nerve. The hair cells of the maculae are covered by a mucous or gelatinous substance which contains calcareous bodies known as otoliths. When the head is in the normal position, the macula of the utricle is in an approximately horizontal position, with the otoliths lying on the hair cells. The macula of the saccule is placed on the lateral wall almost vertically, but slightly inclined so that the planes of the maculae of both sides are not parallel but form an angle opening downward and backward. When the head is in

[1] Usually the direction of nystagmus is given as that of the rapid component; *e.g.*, when the head rotates from left to right, the eyes first move slowly toward the left, then quickly to the right; this is known as nystagmus toward the right.

[1] BARD, *loc. cit.*

[2] WOOLSEY, C. N., H. T. CHANG, and P. BARD, *Federation Proc.*, **6**, 230, 1947.

[3] GAY, N. R., and E. GELLHORN, *Proc. Soc. Exper. Biol. & Med.*, **70**, 711, 1949.

the normal position, the otoliths are placed laterally on the hair cells, embedded in the mucous substance that covers them.

When the head is flexed or extended, the position of the otoliths in relation to the hair cells changes in the utricle. Lateral inclination of the head has the same effect on the otoliths of the saccule. Displacement of the body in a horizontal plane does not change the relative position of the otoliths in the utricle or the saccule. According to Magnus the otoliths stimulate the hair cells by pulling on them.

There are three semicircular canals in each labyrinth (Fig. 476). The external or horizontal canal has its convexity directed laterally and the ampulla forward; it is slightly inclined downward and backward, so that it forms an angle of 30° with the horizontal plane when the head is in the normal position. The anterior canals are placed vertically and in the same plane as the posterior canals on the opposite side; these planes form an angle of 45° with the mesial plane of the body (Fig. 477). The anterior canals have the ampulla forward and the posterior canal backward. The anterior end of the posterior canal is joined to the posterior end of the anterior canal, and both have a common opening in the utricle. The crista acustica lying within the ampulla is the receptor organ innervated by branches of the vestibular nerve. The crista is formed by cells with long cilia or hairs and supporting cells. The hairs are embedded in a gelatinous substance covering the crista, known as the cupula, which occupies the whole width of the canal. The lumen of the canal is filled with fluid called endolymph.

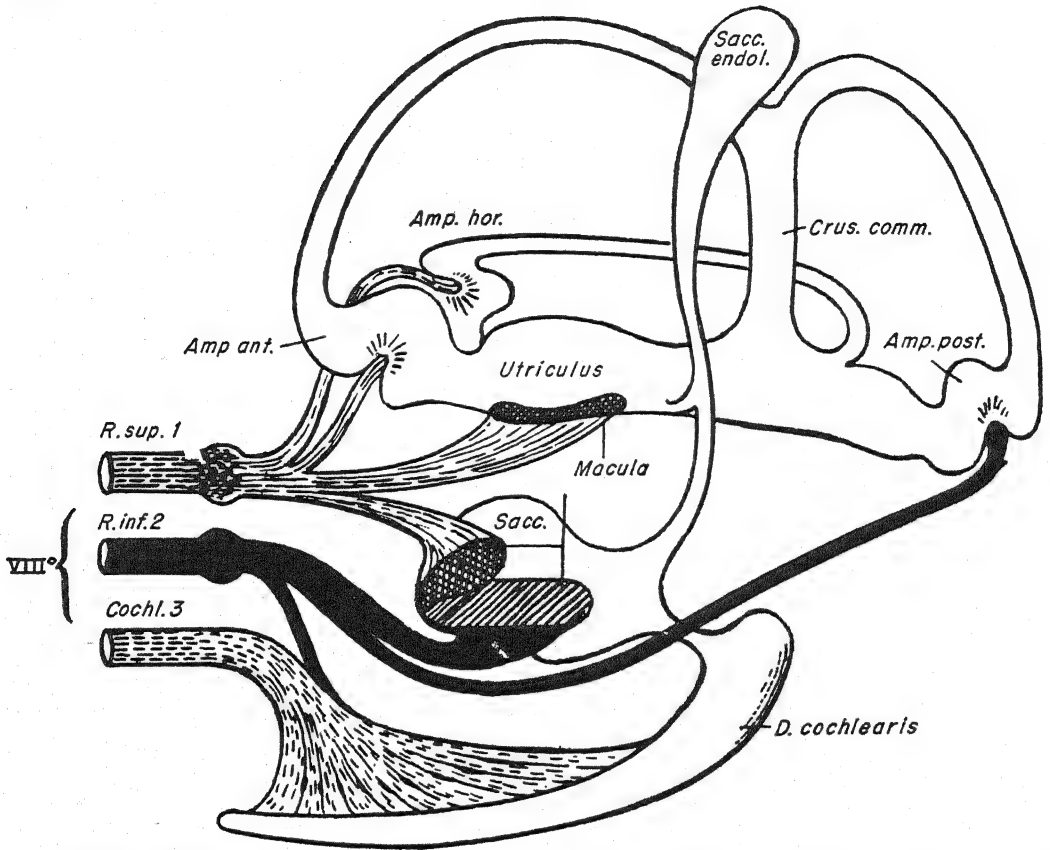



FIG. 476. Diagram of labyrinthine membranous structures and nerves. (*De Burlet, H. M., Anat. Anz., vol. 58, p. 26, 1924.*)

The functions of the labyrinth have been analyzed by observation of the effects of removing one or both labyrinths or of suppressing their activity by cocainization, and by stimulation of the different parts of the organ.

Unilateral labyrinthectomy. Destruction of one labyrinth is followed by asymmetrical reactions due to (*a*) the absence of impulses that

arise normally in the destroyed organ; (*b*) irritation of the vestibular nerve by operative trauma; (*c*) predominance of reflexes initiated in the intact labyrinth.

Ocular symptoms. The eye on the side of the lesion is deviated downward and outward; the eye on the opposite side is deviated upward and outward. There is nystagmus, with the fast component toward the normal side; this is due to irritation of the vestibular nerve and is abolished by cocainization.

Position of the head. The head is rotated so that the face is directed upward and toward the normal side. In some cases there is "head nystagmus."

Extensor tonus. Both limbs on the side corresponding to the lesion are flexed, owing to a decrease in extensor tonus. On the normal side, tonus is increased and the limbs are extended.

Abnormal motility. The animal falls toward the side corresponding to the lesion, owing to the lower extensor tonus. In an effort to compensate this and achieve the normal position, it rolls like a barrel. Ataxia (labyrinthine ataxia) is observed when the animal walks.

The effects are fairly rapidly compensated, and subsequent destruction of the other labyrinth causes reactions similar to those observed after the first operation, but on the opposite side, if the vestibular nuclei are intact on both sides.

Bilateral labyrinthectomy. A double labyrinthine syndrome is observed. The deviation of the eyes and head varies according to which side is more damaged or irritated. Extensor tonus is decreased on both sides, and there is complete loss of postural labyrinthine reflexes. The subject adopts a peculiar stance with widely separated legs, and locomotor ataxia is marked.

Eye and head reactions after unilateral labyrinthectomy are due directly to suppression of one labyrinth with predominance of the other. The effects in the trunk and limbs are caused in part by the abnormal position of the head, and they diminish when the head is rotated back to the normal position. Labyrinthectomy in fishes, birds, and certain mammals (*e.g.*, the rabbit) is followed by severe disturbances; in cats and dogs these disturbances are less severe, and their importance diminishes as encephalization progresses. In primates, especially in anthropoids and man, labyrinthectomy provokes milder and less durable symptoms than in lower animals. Section of the eighth nerve in man is followed by only slight and transitory labyrinthine symptoms.

Symptoms due to irritation of the vestibular nerve soon disappear. On the other hand, labyrinthine reactions are permanently lost after labyrinthectomy, a fact that can be demonstrated by means of several tests. Nevertheless disturbances due to the absence of these reactions are adequately compensated by visual, proprioceptive, and tactile reflexes.

ANTERIOR

POSTERIOR

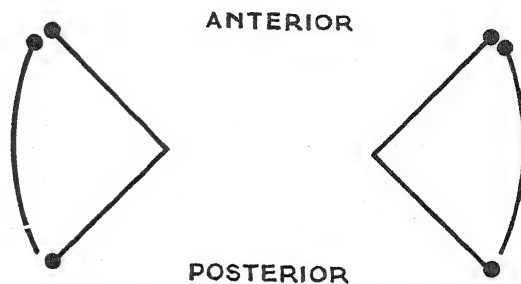



Fig. 477. Diagram of position of semicircular canals.

Functional localization in the labyrinth. The functions of different parts of the labyrinth can be analyzed by destroying or stimulating one or another of them. Thus each one or all of the semicircular canals can be destroyed (Flourens) or stimulated (Ewald). Centrifugation at 1,000 to 2,000 rev./min. detaches the otoliths from the maculae of the saccule and utricle but does not damage the semicircular canals in guinea pigs.

Static labyrinthine reactions arise in the utricle, statokinetic reactions in the semicircular canals. Destruction of the utricle therefore does not disturb statokinetic reactions in response to acceleration, but suppresses almost completely static reactions in response to changes in the position of the head. Destruction of one utricle in the frog[1] is followed by loss of extensor tonus on the side of the lesion; the head and body are bent toward that side and the contralateral limbs are extended. If the animal is stimulated, it progresses in a circle toward the damaged side. Bilateral destruction of the utricle is followed by loss of extensor tonus and severe disturbances in postural reflexes; some of the righting reactions can, however, still be performed. The

[1] Tait, J., and W. J. McNally, *Am. J. Physiol.*, **75**, 155, 1925; *Quart. J. Exper. Physiol.*, **23**, 147, 1933.

utricle therefore plays an important, though not exclusive, part in static labyrinthine reflexes.

The saccule can be destroyed on both sides without disturbing labyrinthine reflexes, even in the rabbit, an animal in which these reflexes are highly developed.[1] The functions of the saccule are not well known; apparently it can be stimulated by vibrations of low frequency.

Functions of the semicircular canals. The semicircular canals can be denervated without provoking disturbances in posture, or abnormal movements, but their destruction is followed by the loss of statokinetic reactions such as the compensatory movements that otherwise occur during rotation.

Ewald stimulated each canal separately by increasing or decreasing the pressure of the endolymph in the following way: A small metal cylinder was cemented over a hole made in the wall of the canal. The narrow end of the canal was plugged, so that changes in pressure, brought about by moving a piston fitted into the cylinder, were exerted on the ampulla. An increase in pressure in the horizontal canals caused the head and eyes to move toward the opposite side; decompression caused a weaker movement in the reverse direction. Compression and decompression of the vertical canals cause similar effects (*i.e.*, deviation of the head and eyes) in the plane of the stimulated canal, but decompression is a more potent stimulus than an increase in pressure. The slow component of nystagmus is of vestibular origin; the rapid component is a compensatory reaction.

The effects of rotation can be easily understood in the light of Ewald's experiments. When the head is rotated with sufficient speed, the endolymph, owing to inertia, is at first displaced at a lower rate than the bony structures, so that it exerts pressure on one end of the canal while the opposite end is decompressed. For example, if the head is rotated in the horizontal plane, from left to right, the endolymph exerts pressure on the ampulla of the right horizontal canal, and the ampulla of the left horizontal canal is decompressed. The eyes move slowly toward the left and are then rapidly brought back toward the right (nystagmus). If rotation continues at a constant rate, the endolymph moves at the same speed as the bony structures; there are no longer differences in the pressure on the ampullas, and nystagmus ceases. When rotation stops, nystagmus commences again (postrotatory nystagmus), but in the opposite direction. The endolymph, owing to inertia, now exerts pressure on the ampulla of the left horizontal semicircular canal, and the ampulla of the right horizontal canal is decompressed.

The movements of the endolymph bend the cupula in the direction in which pressure is exerted, and the "hairs" of the receptor cells are pulled. Dohlman[1] visualized the movements of the cupula by introducing a drop of oil into the canal. He was thus able to see that the cupula bends during the phase of acceleration in rotatory movements and coincides with nystagmus. When the constant phase is established, the cupula returns to the "resting" position and nystagmus ceases. When rotation stops, the same coincidence can be observed between the beginning and ending of postrotatory nystagmus and the bending and straightening of the cupula.

After rotation ceases, there is a sensation of dizziness. The subject feels that rotation still continues, and he sees the objects in the environment rotating in the same direction in which rotation has been taking place. For example, after rotation from left to right, the objects appear to be moving from left to right, *i.e.*, in the same direction as the slow component of the postrotatory nystagmus. Unnecessary compensatory reactions for this apparent movement are performed, and the subject may fall down.

Stimulation of the semicircular canals provokes nystagmus in the plane of that canal (Flourens's law). Therefore to provoke vertical nystagmus the head must rotate in the sagittal plane. This can be performed by bending the head sideways at an angle of 90° and rotating the body in the horizontal plane. Rotatory nystagmus is provoked by flexing or extending the head 90° and rotating in the horizontal plane. Postrotatory vertigo, nystagmus, and compensatory reactions are also observed, and the body falls sideways, forward, or backward according to the position at which the head was held during rotation. For example, if the head is inclined toward the right and the body is rotated from left to right, there is a vertical postrotatory nystagmus with the rapid component downward. The subject feels he is falling forward and performs compensatory reactions

[1] DEKLEIJN, A., and C. J. R. VERSTEEGH, *Pflüger's Arch. f. d. ges. Physiol.*, **232**, 454, 1933.

[1] DOHLMAN, G., *Proc. Roy. Soc. Med.*, **28**, 1371, 1935.

which make him fall backward. If the head is inclined toward the left when rotating, postrotatory nystagmus has the rapid component upward, the subject feels he is falling backward and falls forward.

In the frog the semicircular canals are stimulated by rapidly inclining the supporting surface in the plane of the canal.

Sharp inclination forward and to the left causes movements in the endolymph of the left anterior and right posterior vertical canals. The ampulla of the former is decompressed, pressure is exerted on the ampulla of the latter, and there is extension of the left forelimb. By varying the direction in which the supporting surface is moved, extension of each one of the four limbs can be provoked.

Caloric labyrinthine reactions. On irrigating the external auditory meatus with cold or warm water, convection currents are provoked in the endolymph of the semicircular canals and changes in the pressure on the ampulla result, causing the cupula to bend. The canals are stimulated, and vertigo and nystagmus are observed. The direction of the reactions depends on the temperature of the water, the side stimulated, and the position of the head. If the head is extended 60°, the horizontal canal is placed in an approximately vertical plane. Irrigation of the left ear with warm water (40°C.) or of the right ear with cold water (20°C.) produces horizontal nystagmus to the right (rapid component). Heat produces convection currents toward the ampulla, and cold in the reverse direction; therefore an increase in pressure on the right ampulla will cause slow deviation of the head and eyes toward the left, with a rapid return movement to the right. If the semicircular canals are damaged, these reactions are abolished.

Fig. 478. Vestibular centers and paths.

Galvanic labyrinthine reactions. If an electrode is placed on one of the mastoids, with another electrode on a distant part of the body or on the other mastoid, and a galvanic current (2 to 5 milliamperes)

is passed through the body, the eyes are turned toward the anode and a horizontal or rotatory nystagmus, with the rapid component toward the cathode, is observed. The subject feels dizzy, and objects in the environment seem to be moving from the cathode to the anode. The head and body fall toward the anode. The three semicircular canals are stimulated in this experiment. If the canals are damaged, the same results are obtained by stimulation of the vestibular nerve, but a stronger current (10 to 12 milliamperes) must be used.

Barany's test. A normal subject can place a finger on a given spot even when blindfolded. After rotation he can still do so if his eyes are open, because visual sensations guide the finger to the right place; but if he closes his eyes the finger cannot be accurately placed. For example, after rotating toward the right with the head in the normal position, there is postrotatory nystagmus to the left and the finger deviates or past-points to the left. The subject has the sensation that objects are moving from right to left, and in an effort to "catch up" with the apparently moving object, he passes it and places his finger "ahead" of the object. This is not a vestibular reflex, but a compensatory reaction integrated in the cortex.

Vestibular paths and centers. (Fig. 478.) The maculae of the utricle and saccule and the cristae of the semicircular canals are innervated by fibers of the vestibular nerve, *i.e.*, by axons of cells in the vestibular ganglion (of Scarpa). The central processes of these neurons enter the pons between the inferior cerebellar peduncle and the spinal root of the trigeminal nerve. They then divide into a short ascending branch and a long descending branch, which ends in the vestibular nuclei.

There are four vestibular nuclei: (*a*) the medial (dorsal or principal) nucleus; (*b*) the spinal, or descending, nucleus; (*c*) the superior nucleus of Bechterew; (*d*) the lateral nucleus of Deiters.

The secondary vestibular paths are the following:

1. Fibers from Bechterew's nucleus, together with a few fibers from Deiter's nucleus and others coming directly from the vestibular nerve, form the vestibulocerebellar tract, which enters the cerebellum through the inferior cerebellar peduncle (restiform body) and ends in the cortex of the flocculonodular lobe (see Chap. 83).
2. Fibers from the medial, spinal, and superior nuclei form the medial longitudinal fasciculus, the fibers of which end in the nuclei of the oculomotor nerves (third, fourth, and sixth cranial nerves) of the same and opposite sides. Descending fibers end in the motor nuclei of the cervical segments of the spinal cord. Impulses traveling along this bundle are concerned with the reflex control of movements of the head and eyes.
3. Fibers from the lateral vestibular nucleus descend in the anterior funiculus of the medulla and spinal cord, and end on the spinal motor neurons. Impulses conveyed by this tract facilitate myotatic reflexes.
4. Vestibular paths ending in the thalamus have not been satisfactorily demonstrated, although there are reports of vestibular tracts ending in the lateral part of the ventral and in the medial nuclei of the thalamus. There is suggestive but inconclusive evidence of localization of vestibular function in the temporal lobe.[1] Extirpation of the temporal cortex, however, is not followed by signs of vestibular disturbance or alterations in labyrinthine reactions.

BIBLIOGRAPHY

BARD, P., *et al.*, Patterns of Organization of the Nervous System, *Research Publ., A. Nerv. & Ment. Dis.*, **30**, 1952.

CAMIS, M., "The Physiology of the Vestibular Apparatus," translated and annotated by R. S. Creed, Oxford, New York, 1930.

COBB, S., Review on the Tonus of Skeletal Muscle, *Physiol. Rev.*, **5**, 518, 1925.

CREED, R. S., D. DENNY-BROWN, J. C. ECCLES, E. G. T. LIDDELL, and C. S. SHERRINGTON: "Reflex Activity of the Spinal Cord," Oxford, New York, 1932.

DOW, R. S., The Effect of Unilateral and Bilateral Labyrinthectomy in Primates, *Am. J. Physiol.*, **121**, 392, 1938.

FULTON, J. F., Somatic Functions of the Central Nervous System, *Ann. Rev. Physiol.*, **15**, 305, 1953.

LORENTE DE NÓ, R., Researches on Labyrinth Reflexes, *Trans. Am. Otolaryng. Soc.*, **22**, 287, 1932.

MAGNUS, R., "Körperstellung," Springer, Berlin, 1924.

MAGNUS, R., Studies in the Physiology of Posture, *Lancet*, **211**, 531 and 585, 1926.

MAGOUN, H. W., Caudal and Cephalic Influences of Brain Stem Reticular Formation, *Physiol. Rev.*, **30**, 459, 1950.

SHERRINGTON, C. S., "The Integrative Action of the Nervous System," Yale University Press, New Haven, 1926.

[1] SPIEGEL, E. A., *Arch. Neurol. & Psychiat.*, **31**, 469, 1934.

CHAPTER 82

Cortical Integration of Movement. The Pyramidal and Extrapyramidal Motor Systems

THE ACTIVITY OF the nervous system consists in the integration of functional patterns. The two-neuron arc reflex is the simplest integration and takes place in the spinal cord. The complexity of integration increases by the interpolation of internuncial neurons in the reflex arc, as occurs in segmental reflexes (*e.g.*, flexion and crossed-extension reflexes) and in the more complex intersegmental, or "long," spinal reflexes (*e.g.*, the scratch reflex). In the intact animal the spinal level of integration is under the control of higher centers, and even relatively simple spinal synergies may be modified by positive or negative impulses from a higher level, *i.e.*, these impulses may facilitate or inhibit spinal reactions. Outstanding cases of this supraspinal control are the conditions of spinal shock, where facilitation is absent, and decerebrate rigidity, where inhibition is lacking. Beside this effect on spinal reflexes, supraspinal centers integrate patterns of greater complexity than those integrated in the spinal cord. Perfect integration of the whole body requires the activity of the cerebral cortex, *i.e.*, a cortical level of integration, but after the cortex has been removed, even relatively complex synergies are integrated at subcortical levels of function, *e.g.*, the postural reflexes in the somatic sphere and the regulation of blood pressure in the visceral sphere. When the subcortical centers have been destroyed, isolated parts of these synergies can still be carried out, but their complete and harmonious accomplishment is no longer possible.

Effects of total extirpation of the cortex. Cortical integration increases in importance with the development of the nervous system and the process of encephalization. This fact can be demonstrated by experimental removal of the whole cortex, an operation that results in more marked disturbances the higher the position of the subject in the vertebrate scale. In fishes and amphibians, ablation of the cortex does not disturb reflex and automatic movements. The following are the most outstanding deficiencies: (*a*) there are no spontaneous movements, but posture and swimming are normal; (*b*) the animal does not search for food, although food placed in the mouth is swallowed and digested normally.

Decortication produces more marked effects in birds, which can fly for only a very short stretch in a well-lit environment and remain motionless in the dark.

Goltz, in 1881, showed that dogs can live for a long time (18 months) after the removal of the whole cerebral cortex. Phasic and postural reflexes are apparently normal, and the animals can acquire and maintain different postures (see "Centers that coordinate postural reactions," Chap. 81). Automatic movements such as walking are also normally performed. The animals remain quiet unless stimulated to move, but once in motion they continue to move incessantly and appear to be restless. Anticipatory reactions usually initiated in the distance receptors are not observed; harmful agents do not evoke defense reflexes until actual damage has occurred, and then an exaggerated emotional

response may be seen (pseudoaffective state, sham rage; see "The hypothalamus," Chap. 84). The animals appear irritable, and weak stimuli provoke unusually strong reactions. They do not search for food, but when given it is swallowed and digested; there are no outstanding changes in visceral functions. Objects and persons are not recognized; memory is lost, and new happenings are not stored in the mind. All "learned" reactions are forgotten, and the capacity to learn has been lost.

Primates do not survive for long after total decortication. With great care, however, decorticate children have been kept alive for a relatively long time. Even removal of only part of the cortex produces more marked effects than those observed in the cat and dog after removal of the whole cortex.

The following methods have been used in the study of cortical functions: (*a*) stimulation; (*b*) observation of the effects of destruction of limited parts on innate and acquired movements, such as conditioned reflexes and trained movements, and on general behavior; (*c*) the registration of electrical potentials; (*d*) strychninization of limited areas. The correct interpretation of the results obtained by these methods requires the knowledge of cortical structure and of the afferent and efferent pathways.

Fritsch and Hitzig (1870) were the first to demonstrate that stimulation of the frontal cortex produces movements in the limbs of the opposite side. Shortly before, Jackson had seen cases of epilepsy in which convulsions began in a discrete group of muscles, *e.g.*, in the thumb muscles, and from there spread to the neighboring regions and then to the rest of the body, following a typical and constant "march." He deduced that in the cerebral cortex there were centers that controlled isolated movements. In cases where a pathologic cause irritates one of these centers, there is first a focal seizure, and on persistence of the stimulus, excitation spreads to the rest of the cortex. Fritsch and Hitzig's cortical excitation experiments and those of Ferrier on the results of localized extirpation of the cerebral cortex in monkeys gave a solid experimental basis to Jackson's conception, which has since received ample proof.

FIG. 479. Pyramidal and extrapyramidal cortical areas. Lateral view of cerebral hemispheres with principal cytoarchitectural areas. (*After Foerster.*)

The motor cortex is considerably more developed in primates, particularly in the anthropoids and man, than in other species. It is situated in the precentral convolution (area 4). This is the only part of the cortex from which discrete, well-localized movements can be obtained by stimulation, but movements can also be observed when other parts are stimulated. Thus stimulation of the frontal lobe, rostrally to the motor cortex on the premotor area 6, and area 8 (frontal eye field), parietal lobe (areas 3-1-2, 5, and 7), temporal lobe (area 22), occipital lobe (area 19), and parts of the rhinencephalon provokes movements. Some of these movements are suppressed when area 4, or its connections with the cortical center under stimulation, are destroyed. Area 4 sends out the pyramidal tract; for this reason it is known, together with its projection fibers, as the pyramidal motor system. The rest of the excitable cortex and its projection fibers are called the extrapyramidal motor system (Fig. 479).

THE PYRAMIDAL MOTOR SYSTEM

Anatomic description. The motor cortex (area 4) extends from the depth of the sulcus centralis, up its anterior wall, to the lateral aspect of the prefrontal convolution, having a breadth of about 1 cm. on the upper part and tapering down as the sylvian fissure is approached; it also extends along the median surface of the hemisphere down to the sulcus cinguli. Rostrally it is separated from area 6 by Marion Hines's "strip" region (4*s*). The large pyramids of Betz are found in this area; section of the pyramidal tract produces chromatolysis in these cells. The Betz cells are not the only neurons that send out axons to make up the pyramidal tract, since in man there are only 25,000 to 30,000 of them on each side, and approximately 1,000,000 pyramidal fibers, 61 per cent of which are myelinated.[1] Other neurons therefore contribute the majority of the pyramidal fibers. All the myelinated fibers in the pyramidal tract have a cortical origin, as is shown by the fact that hemidecortication in monkeys is followed by the total disappearance of fibers stained by the Weigert technique.[2] The origin of the unmyelinated fibers in the pyramidal tract has not yet been demonstrated.

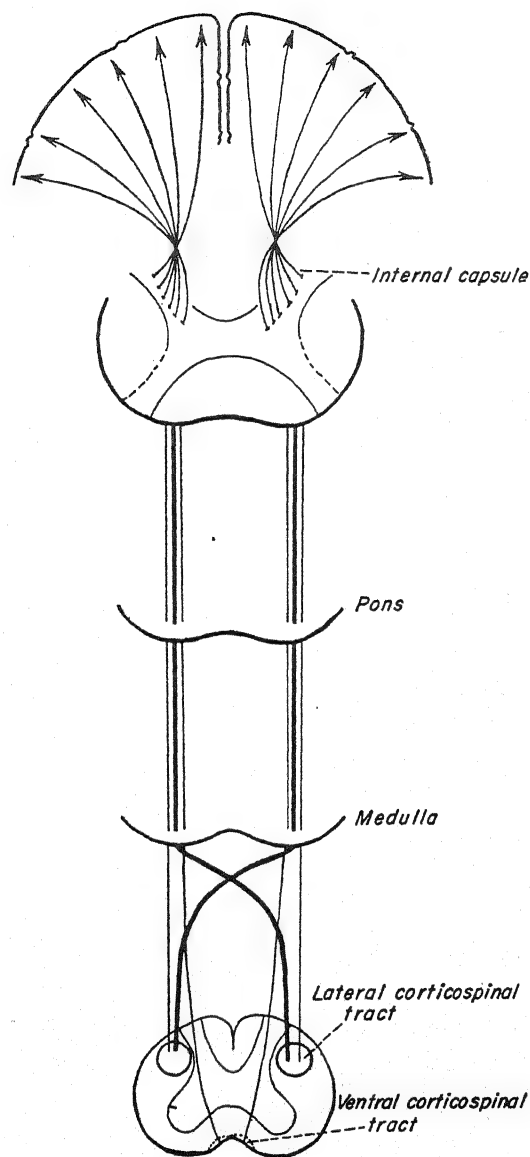



FIG. 480. Pyramidal tract.

Antidromic stimulation of the pyramids in the medulla[3] has shown that the cortical origin of the pyramidal tract is widely spread. The electrical response is maximal in area 4, but almost as great in the postcentral convolution (area 3-1-2), and it extends to area 6 in the frontal lobe and to areas 5 and 7 in the parietal lobe.

The pyramidal tract (Fig. 480) descends along the anterior two-thirds of the posterior limb of the internal capsule, where it is sometimes destroyed owing to vascular lesions. Farther down it occupies the middle three-fifths of the basis pedunculi and then forms the pyramids of the pons and medulla. In the lower part of the medulla the tracts cross to the opposite side (decussation of the pyramids).

There is a certain degree of systematization all along the corticospinal tract. The fibers arising in the median aspect of the cerebral hemisphere and the upper part of the precentral gyrus (area 4*a*) descend along the posterior part of the internal capsule and the lateral part of the pedunculi, pons, medulla, and spinal cord (Fig. 400); they end in the lumbar and sacral segments. Medial to these fibers are those arising in area 4*b*, which end in the lower cervical and thoracic segments. The fibers in the genu of the internal capsule occupy the median part of the tract in the pedunculi. They come from the lower part of the precentral gyrus (area 4*c*) and end in the nuclei of the cranial motor nerves and the upper cervical segments.

The pyramidal fibers in the spinal cord can be divided into three groups: (*a*) the lateral or crossed corticospinal tract, which includes 70 to 85 per cent of all pyramidal fibers in man; (*b*) the ventral or direct

[1] LASSEK, A. M., *Arch. Neurol. & Psychiat.*, **42**, 872, 1939; **44**, 718, 1940.

[2] WELCH, W. K., and M. A. KENNARD, *J. Neurophysiol.*, **7**, 255, 1944.

[3] WOOLSEY, C. N., and H. T. CHANG, *Federation Proc.*, **6**, 230, 1947.

corticospinal tract, which is not prominent, is seldom found below the cervical segments, and is made up of fibers coming principally from area 4*c*; (*c*) direct fibers which do not take part in the decussation of the pyramids, but descend in the ipsilateral corticospinal tract; they are 10 per cent of all the pyramidal fibers in the chimpanzee. Almost all the fibers (80 to 90 per cent) end on internuncial neurons and not on spinal motor neurons (Fig. 481).[1] Part of the corticospinal projection (20 to 25 per cent) is therefore not crossed, but ends on ipsilateral internuncial or motor neurons (Fig. 480).

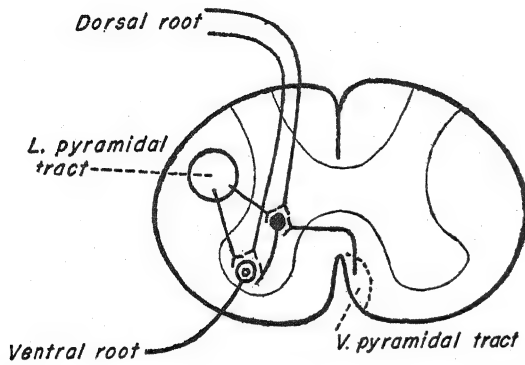



Fig. 481. Ending of pyramidal tract in spinal cord.

Brodal and Walberg[2] have observed ascending degeneration after section of the ventral or lateral corticospinal tracts in cats. These degenerating fibers may be seen when the section is as low as the fifth lumbar segment of the cord, but the majority arise in the cervical cord. They have been traced up to the pyramids and the cerebral peduncles; in the pontine gray matter, terminal degeneration was seen, and some fibers end in the motor cortex and a larger number in the sensory cortex. These ascending fibers in the corticospinal tracts form about 4 per cent of the total number of fibers. Kaada[3] has recorded ascending impulses by means of electrodes implanted in the pyramids when the skin of the limbs was stimulated (especially the contralateral forelimb). The significance of these fibers is not yet known.

Stimulation of the motor cortex (area 4). Stimulation of area 4 produces discrete movements, which can be analyzed by means of electromyographic records. In cats and monkeys a very weak stimulus applied to the cortex produces relaxation in muscles tonically contracted. A slightly stronger stimulus provokes movements similar to those observed in phasic reflexes; there is contraction of the protagonists and relaxation of the antagonists according to the principle of reciprocal innervation. A still stronger stimulus is followed by simultaneous contraction of the protagonists and antagonists of phasic movements (*e.g.*, flexors and extensors) and the limb is fixed.[1]

If the cortex is stimulated in conditions opposing intracortical conduction so that excitation does not spread, the response obtained may be limited to a single muscle or part of a group of muscles. A neuron pool, surrounded by a fringe that overlaps the fringes of neighboring pools, corresponds to each muscle or group of muscles.[2]

The optimum intensity of stimulation varies for different centers, but the duration of the stimulus is always relatively long. Thus the chronaxie for the center from which movements of the fingers are provoked is 1 msec.; for flexion of the hip it is 2 msec.[3] The optimum frequency for repeated stimulation also varies in different centers. Using alternating current, for area 4 it is usually below 25 c.p.s. Thus different cranial motor nuclei are excited when the face area of the cortex is stimulated at different frequencies.[4]

Further evidence of this fact has been obtained by stimulation of the bulbar pyramids in cats and monkeys. Motor neurons respond only to repetitive bombardment of internuncial neurons, and the duration of the threshold pyramidal volley varies for each muscle. Large, fast-conducting, low-threshold (short-duration) fibers end on motor neurons corresponding to the face, or fingers; smaller, slower fibers with a higher threshold (long duration) activate motor neurons of proximal limb muscles.[5]

The excitability of the cortex, as well as that of other parts of the nervous system, is modified by several factors. Following the application of an effective stimulus there is a period of depres-

[1] Hoff, E. C., *Proc. Roy. Soc., London, s.B.*, **111**, 226, 1932.
[2] Quoted by Fulton, J. F., *Ann. Rev. Physiol.*, **15**, 305, 1953.
[3] Quoted by Fulton, *loc. cit.*

[1] Gellhorn, E., and J. F. Bosma, *Federation Proc.*, **5**, 32, 1946.
[2] Chang, H. T., T. C. Ruch, and A. A. Ward, Jr., *J. Neurophysiol.*, **10**, 39, 1947.
[3] Wyss, O. A. M., and S. Obrador, *Am. J. Physiol.*, **120**, 42, 1939.
[4] Bartley, P., and G. von Bonin, *Tr. Am. Neurol. A.*, p. 89, 1946.
[5] Brookhart, J. M., *Research Publ., A. Nerv. & Ment. Dis.*, **30**, 157, 1952.

sion, called "extinction,"[1] which reaches its maximum in 14 sec. and disappears in 30 to 40 sec. This phase of hypoexcitability is similar to that observed in the spinal cord and peripheral nerves, and it also coincides with a positive potential. Alkalosis, such as is produced by forced breathing, increases cortical excitability. Acidosis has mainly a depressing effect.

The different parts of the body are focally represented in the precentral convolution (Fig. 405). There is the same general distribution in all vertebrates. The parts innervated by the caudal spinal segments are represented in the cortex near the mid-line, and the more rostral segments more laterally. Thus the sacral segments (tail, perineum) have their representation in the median aspect of the hemisphere and the upper part of the precentral gyrus, and the hind limb in the adjacent parts of the gyrus, lateral to the former. Next to these are the foci corresponding to the trunk. The forelimb is represented in area 4*b*. The head and face are represented in the most lateral part of the convolution (area 4*c*), with the peculiarity that the foci pertaining to the face are lateral to those of the head, an inversion of the general order similar to that observed in the sensory centers of the postcentral convolution. This representation is principally, but not exclusively, contralateral. The extent of the cortical representation is in direct relation to the complexity and fineness of movement. Thus movements of the thumb and of the tongue and lips occupy a large surface, while those of the forehead and scalp occupy only a small one (Fig. 482). In man stimulation of area 4*c* provokes grunts and cries.

The following facts show that the movements produced by stimulation of the motor cortex are originated in the Betz cells of layer V:

1. In the newborn human infant and the monkey under one month old, area 4 is not excitable, because the cortex is not yet completely developed and the axons of the Betz cells are not myelinated.
2. The different layers of the cortex can be progressively destroyed by coagulation provoked by the application of heat at a suitable temperature (70 or 80°C.) for a limited time (3 to 5 sec). Area 4 continues to respond after the outer layers have been destroyed, but ceases to respond when layer V is destroyed. The α rhythm of the electroencephalogram also remains undisturbed as long as layer V is intact.
3. Section of the pyramidal tract produces chromatolysis in the Betz cells and suppresses the localized response to excitation of area 4. Movements provoked by cortical stimulation are of the extrapyramidal pattern (see pages 1041–1042).

[1] Dusser de Barenne, J. G., and W. S. McCulloch, *J. Neurophysiol.*, 2, 319, 1939.

a

b

c

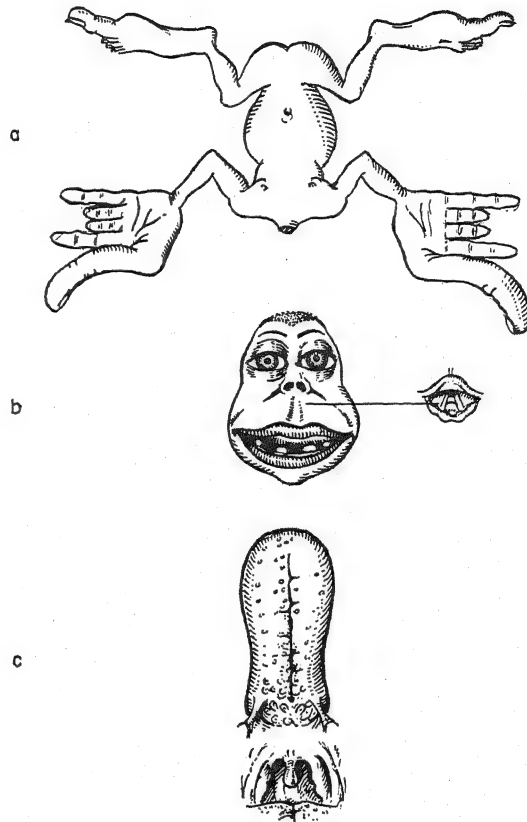



Fig. 482 Homunculus illustrating position and extent of cortical representation in motor cortex. Part *c* shows the throat. (*Penfield, W. G., and E. Boldrey, Brain, vol. 60, p. 389, 1937.*)

If a particular center is persistently stimulated, first a localized response is provoked, then excitation spreads to the adjacent foci and can extend to the whole motor area, following a "march" in the same order as in jacksonian epilepsy. Convulsions last for 3 to 4 min., alternating with periods lasting 5 to 6 min. during which there is no response in spite of the persistent stimulation. With a sufficiently strong stimulus, the ipsilateral limbs also enter into convulsions, probably because of the uncrossed fibers of the pyramidal tracts. Stimulation of

the motor cortex after contralateral semisection of the spinal cord is followed by isolated movements on the ipsilateral side; the impulses travel along the 25 per cent uncrossed contingent of the corticospinal tract.

Facilitation and inhibition take place in the motor cortex, as in other nerve centers. Repeated subliminal stimuli provoke a response, but leave a subsequent period of depression. Stimulation of other parts of the cortex, such as area 3-1-2 in the parietal lobe and the premotor area 6, facilitates stimulation of area 4. If the connections between these areas and the motor cortex are cut, this effect is suppressed. There are also areas of the cortex from which inhibition of area 4 can be obtained. The first of these inhibitory areas to be described was the "strip" (area 4) discovered by Hines[1] in the monkey; it is situated immediately rostral to area 4, between this area and area 6. Other inhibitory areas were described later in the frontal lobe (area 8*s*), in the parietal lobe (area 2*s*), and in the occipital lobe (area 19*s*). Stimulation of the inhibitory areas suppresses movements originated in area 4, prevents facilitation, and prolongs the positive after-potentials.[2] Localized strychninization of the inhibitory areas enhances their negative influence on the motor cortex.

Facilitation and inhibition of cortical movements can also be obtained by stimulation of subcortical centers. Thus stimulation of the hypothalamus facilitates movements originated in the motor cortex. These centers do not act on the cortex but on the spinal centers, because facilitation of movement is also observed after the cortex has been removed. There are a series of facilitatory centers extending backward from the diencephalon to the bulboreticular substance.[3]

Cortical inhibition induced by subcortical centers is a prominent feature of cortical function which will be discussed later in connection with the basal ganglia.

The second somatic effector area.[4] A second somatic motor area has been found lying in front of, and in part overlapping, the second somatic receptor area (see Chap. 74). It is situated in the wall of the frontoparietal operculum and the posterior part of the insula. The different parts of the body are represented in an order opposite to the one in which they are represented in area 4; the face area is anterosuperior and the leg area posteroinferior to the arm area, which is the largest and is placed between the other two. Stimulation provokes movements in the other side of the body, especially in the distal parts of the limbs. These movements can be elicited even after area 4 has been removed. Extirpation of the second somatic effector area is not followed by lasting disturbances in motor functions.

Extirpation of the motor cortex. Ablation of the motor cortex (area 4) is followed by flaccid paralysis on the opposite side; *e.g.*, if area 4*b* on the right side is destroyed, the left forelimb is paralyzed. Immediately after the operation the limb remains motionless and flaccid, but after a time there is a partial recovery of movement. The effects are more marked and lasting, and the final deficit is greater, as encephalization is more pronounced and a larger area is removed. The final deficit is greater if area 3-1-2 of the postcentral lobe is also removed. After 1 to 3 days in monkeys and anthropoids (chimpanzee), movements of the proximal part of the limb (shoulder, hip) can be observed. Motility increases gradually, and after 3 to 4 weeks in the monkey and 8 to 12 weeks in the chimpanzee, some of the movements of the hand have been recovered. In the upper limb, flexor muscles show greater recovery, and in the lower limb, extensor muscles. Recovery, however, is never complete; the limb is clumsy, and fine movements and those acquired by learning are lost. This deficiency is not due to loss of memory but to the motor deficit of the limb; the animal knows the complicated movements it had learned, but cannot carry them out and tries to compensate for the defect with its teeth or another limb. It can learn new movements, but its performance is limited by paralysis. In man the final deficit is even more considerable. Certain drugs, such as doryl, strychnine, and thiamine, have a beneficial effect on recovery after ablation of areas 4 and 6 in the monkey; others, *e.g.*, phenobarbital, retard recovery.[1]

Reflexes are at first depressed; flaccidity is

[1] HINES, MARION, *Am. J. Physiol.*, **116**, 76, 1936.

[2] DUSSER DE BARENNE, J. G., H. S. GAROL, and W. S. MCCULLOCH, *J. Neurophysiol.*, **4**, 324, 1944.

[3] RHINES, R., and H. W. MAGOUN, *J. Neurophysiol.*, **9**, 219, 1946.

[4] SUGAR, O., J. G. CHUSID, and J. D. FRENCH, *J. Neuropath. & Exper. Neurol.*, **7**, 182, 1948.

[1] WATSON, C. W., and M. A. KENNARD, *J. Neurophysiol.*, **8**, 24, 1945.

complete, owing to the absence of the myotatic reflex. After a time reflexes recover their normal strength, and some of them may be increased; there is transitory spasticity of the fingers. Postural tonus and reflexes also improve, but recovery is not complete. If the lesion encroaches on Hines's "strip" area (4*s*), there is a certain degree of spasticity, which is considerably greater if area 6 is also destroyed. Man and the anthropoids have a positive Babinski sign, which appears whenever the pyramidal tract is damaged. This consists in disturbance of the flexion reflex of the lower limb. Normally the toes are flexed when the plantar surface is stimulated, but if there is a lesion in the pyramidal tract there is extension, especially of the great toe, and in some cases there is "fanning" of the toes, which is a sign of extrapyramidal encroachment.

After a time there is some muscular atrophy. There are also signs of autonomic disturbance, especially in vasomotor reactions (see "Cortical integration of visceral functions," Chap. 84).

Section of the pyramids in the medulla is also followed by a flaccid paralysis, without any signs of spasticity. If the corresponding area 4 is subsequently destroyed, paralysis increases and a certain degree of spasticity develops; this is especially noticeable in the resistance offered to passive movements. If, on the contrary, area 4 is stimulated after section of the pyramids, an inhibitory effect is observed, which is transmitted along extrapyramidal paths.

The functions of the pyramidal system can be considered from two points of view, according to Tower:[1]

1. The localization of function, which refers to its spatial organization. There are centers of representation of discrete, isolated movements, which are more developed for the more complicated and delicate movements. These centers are distributed in a roughly segmental pattern, and are predominantly contralateral. The impulses travel along the corticospinal tract, which also has a certain systematization.
2. The pyramidal tract functions in two phases. There is a continuous discharge of impulses, which facilitates spinal reflexes and keeps their threshold low. When these impulses are lacking there is flaccidity. Spinal shock in primates is due principally to the suppression of this corticospinal bombardment of the spinal internuncial and motor neurons. The "tonic" influence of the cortex extends to somatic and visceral reflexes. It varies from time to time; it is at its lowest during sleep and increases on awakening. Besides these continuous diffuse discharges, there are others limited to groups of neurons which produce isolated, well-defined movements and are necessary to perform delicate and complicated movement patterns. There are also inhibitory neurons in area 4, the impulses of which are not transmitted along the pyramidal tract, but along extrapyramidal paths.

[1] Tower, S., *Brain*, **63**, 36, 1940.

The pyramidal system is usually considered as the path of voluntary movements. This is so in respect to isolated movements, but even after destruction of the motor cortex some voluntary patterns can still be performed. Therefore it is not the only system carrying voluntary nervous impulses.

THE EXTRAPYRAMIDAL MOTOR SYSTEM

The cortical centers of the extrapyramidal motor system receive fibers from the thalamus and send out fibers to the nuclei of the diencephalon and, through these, to the pontobulbar and spinal motor centers. Its pathways are all short; none is as long as the spinothalamic and pyramidal tracts. There are two main paths: (*a*) the corticostriatomesencephalic projection, which integrates posture and automatic movements; (*b*) the corticopontocerebellar projection, which integrates voluntary movements. Both these projections act on the efferent side of the great reflex arc that goes from the receptors to the cerebral cortex and from the cortex to the effectors.

Stimulation of the extrapyramidal cortex. Stimulation of several parts of the cortex outside area 4 provokes movements. Some of these movements are mediated through the motor cortex, and they are suppressed by ablation of area 4, or by cutting the connections with this area. Others are not dependent on the integrity of the pyramidal system; they are mediated through extrapyramidal paths. The threshold of the extrapyramidal cortex is considerably higher than that of the motor cortex.

Stimulation of area 6 (premotor cortex) produces two types of movement:

1. Discrete circumscribed movements, not so finely limited as those produced by stimulation of area 4. They are suppressed by severing the connections between areas 4 and 6 or by ablation of area 4.
2. Complex movement patterns involving the whole limb (sometimes several limbs), which are carried out slowly. They are not dependent on the integrity of the pyramidal system.

Well-localized, discrete movements are obtained on stimulation of the caudal parts ($6a\alpha$), seldom from the more rostral regions ($6a\beta$). "Adversive" movements (turning of the head and eyes toward the opposite side) are provoked by stimulating area $6a\beta$. Stimulation of the more lateral part of the premotor cortex (area *6b*) situated rostrally to area *4c*, produces movements of the tongue and lips, chewing and swallowing, and in man grunts and cries. These movements are not mediated by the motor cortex, and they are inhibited by stimulation of area 8γ. Stimulation of the lateral part of the premotor cortex spreads to the orbital surface of the frontal lobe, and produces the same effects as stimulation of area 13, *i.e.*, inhibition of respiratory movements and of gastrointestinal motility.

In the posterior part of the second frontal convolution, rostral to $6a\alpha$ and lateral to $6a\beta$, there is an area ($8\alpha\beta\delta$) known as the frontal eye field. Stimulation of this region provokes conjugate movements of the eyes to the opposite side, opening of the eyelids, and sometimes dilatation of the pupil. Closure of the eyes can be obtained by stimulating a point situated more laterally.

Stimulation of the postcentral convolution (area 3-1-2) with a strong stimulus occasionally provokes discrete movements, such as are represented in the neighboring focus of the precentral convolution. Subliminal stimulation of area 4 facilitates this effect, and stimulation of the postcentral gyrus facilitates response to stimulation of the motor cortex. All these effects are suppressed if the connections between the precentral and postcentral convolutions are cut.

Stimulation of areas 17 and 18 in the occipital lobe, and of area 19, provokes movements of the eyes, downward when the upper part, and upward when the lower part, of the occipital lobe is stimulated (occipital eye fields). Electrical stimulation of the rostral part of the cingular gyrus in the monkey (area 24) provokes complex somatic and autonomic responses. There is vocalization (cries, grunts), facial and other movements related to the expression of emotion, opening of the eyelids, dilatation of the pupil, erection of hairs, depressed respiratory rhythm (with occasionally apnea of short duration), and cardiovascular reactions.[1]

Inhibitory areas. Inhibition of area 4, and of the whole cortex, can be obtained by stimulating several "inhibitory" strips. Besides the strip region between areas 4 and 6, there are those already mentioned in areas 8, 2, and 19. Strychninization of these strips diminishes the electrical activity of the whole cortex. Strychninization of area 6, on the contrary, increases electrical activity over almost the whole cortex.

There is a powerful inhibitory area in the rostral end of the cingular gyrus.[2]

Electrical stimulation of this area (Brodmann's area 24) causes cessation of spontaneous movements. The animal closes its eyes and apparently remains asleep for several minutes after stimulation has ended. Inhibitory effects following stimulation of the rhinencephalon are not limited to area 24; stimulation of the posterior orbital gyrus around the olfactory tubercle, the anterior insula, anterior hippocampal gyrus, and neighboring temporal cortex suppresses spontaneous movements in monkeys and cats. Stronger stimuli may evoke slow tonic movements, often of the adversive type. Movements evoked by stimulation of the motor cortex, and in a lesser degree spinal reflex movements, may be facilitated, or stimulation may be followed by inhibition, or inhibitory effects alone are observed; this occurs more frequently when the areas around the olfactory tubercle and the cortex near the genu of the corpus callosum are stimulated. These areas exert their effects independently of the neocortical motor areas; removal of the motor cortex and even of a large part of the neocortex does not influence the response to stimulation of the anterior rhinencephalon. Inhibition and facilitation are proba-

[1] Smith, W. K., *J. Neurophysiol.*, 8, 241, 1945.

[2] Bailey, P., G. von Bonin, E. W. Davis, H. W. Garol, W. S. McCulloch, R. Roseman, and A. Silveria, *J. Neurophysiol.*, 7, 51, 1944.

bly mediated through the diencephalic and mesencephalic reticular systems.[1]

After removal of the whole cortex (neocortex and archicortex) the animals show exaggerated emotional reactions (sham rage; see "The Hypothalamus," Chap. 85). If the neocortex is removed but the olfactory cortex is intact, the animals remain quiet and placid. Removal of the olfactory cortex is followed by exaggerated emotional reactions.[2]

The existence of circuits between these cortical inhibitory areas and the striatum has been demonstrated by the strychninization method (see "Corticostriatal connections," page 1046), but areas 4*s*, 8*s*, 2*s*, 19*s*, and 24 also discharge inhibitory impulses into the pontobulbar reticular formation, and possibly by a series of relays, down to the spinal centers.[3]

Extirpation of the extrapyramidal cortex. Unilateral ablation of the premotor area (6*a*) produces the following disturbances on the opposite side:

1. Clumsiness in movements, especially in those acquired by learning. The animals have more difficulty in the performance of complex movement patterns than in performing discrete movements. A long period of training is necessary to reacquire the lost skill.
2. Spasticity, which diminishes as time passes.
3. Increased reflex response, which is more marked if area 4 has been previously removed. There is "fanning" of the toes on plantar stimulation, and an exaggerated grasp reflex (slow and strong flexion of the fingers or toes on palmar or plantar stimulation).

Spasticity and hyperreflexia are release phenomena due to the liberation of lower centers from cortical influence. Inhibitory neurons are not restricted to area 6 but are found in many other parts of the cortex. Ablation of area 4 produces a short-lasting spasticity limited to the distal part of the extremities. If the strip area is also destroyed, spasticity of the whole limb is observed. Permanent spasticity of an even greater degree is produced if area 6 is also included in the operation. The parietal lobe has inhibitory neurons; there is no spasticity after area 3-1-2 has been destroyed, but it does occur if area 4 is removed at the same time. The inhibitory influence of the frontal lobe is demonstrated by the increase in spasticity that follows its removal subsequent to destruction of areas 4 and 6. A maximal degree of spasticity is observed on complete decortication, in which case it is similar to decerebrate rigidity but less marked; muscles that counteract the effect of gravity are those most affected.

Other features observed after removal of area 6*a*, which are due to loss of inhibition, are (*a*) groping, which consists in repeated reaching movements preparatory to grasping; (*b*) perseveration, *i.e.*, movements are repeated and postures are held long after their usefulness has passed; *e.g.*, chewing continues when there is no longer food in the mouth.

Spasticity, rigidity, and hyperreflexia are due to release of subcortical centers; groping and perseveration are due to release of cortical centers and are not observed when the appropriate cortical areas have been destroyed; *e.g.*, groping is mainly of visual origin and ceases when vision is abolished.

Unilateral removal of the frontal eye fields ($8\alpha\beta\delta$) is followed by twisting of the head toward the operated side and forced circular movements of the body in the same direction. For some time, conjugate movements of the eyes toward the opposite side cannot be performed. Bilateral ablation of these areas is followed by transient visual agnosia; light perception is retained, but the nature and significance of objects are not recognized. When the lesion is restricted to one side, there is hemiagnosia; objects are not recognized while they occupy the visual field on the opposite side but are quickly identified when they come into the visual field on the side of the lesion. The disturbance is apparently due to lack of correlation of visual impulses integrated in the occipital eye fields (areas 18 and 19) with extraocular proprioceptive impulses integrated in the frontal eye fields (area 8).

Frontal association areas (9, 10, 11, 12, 13, 45, 46, and 47). The anterior pole of the frontal lobe is considerably developed in primates, especially in the anthropoids and in man. It contains extrapyramidal inhibitory neurons, but

[1] KAADA, B. R., *Acta physiol. Scandinav.*, **24**, Suppl. 83, 1951.

[2] BARD, P., and V. B. MOUNTCASTLE, *Science*, **107**, 457, 1948.

[3] MCCULLOCH, W. S., and E. HENNEMAN, *Federation Proc.*, **7**, 79, 1948.

the main concern of the anterior pole of the frontal lobe is nonspecific intracortical association, and for this reason it is called the frontal association area. This is a "silent" area; faradic stimulation does not produce movement nor sensory phenomena, and ablation is not followed by any disturbance in reflex, postural, or voluntary movements, or any sensory defect.

Frontal lobectomy must be bilateral to produce marked effects; few results, if any, and these transient, are observed after unilateral damage. The disturbances registered are not of movement or sensation but of behavior.

Hyperactivity is one of these disturbances. After a period of depression, during which the animals remain motionless and apathetic, they commence to have fits of hyperactivity, which increase as time passes. Stereotyped, aimless movements, such as pacing back and forth in the cage, are repeated for long periods. There is not much variety in the movement patterns, and perseveration is an outstanding feature. Hyperactivity is maximal and can assume maniacal proportions when area 13 on the orbital surface is destroyed.[1] There is also increased gastrointestinal motility, accompanied by hyperphagia, and the subject eats and drinks excessively.

Distractibility is another typical disturbance; attention cannot be kept fixed on an object, because any intercurrent stimulus distracts the animal. For example, if it is given a grape, it will be carried to the mouth but dropped immediately if a second grape is given, and this one also will be dropped if a third one is offered, and so on until the animal is surrounded with fruit that it has let fall without having eaten a single one.

Immediate memory is lost, events cannot be fixed in the mind, and therefore recent experience cannot influence behavior. The animals "live in a perpetual present," without remembering what has just occurred or anticipating the immediate future. To demonstrate this deficiency, several methods have been used which consist in testing the ability to respond after a delay. For example, an animal is shown how food is placed below one of two cups, which are then hidden for a short time before the animal is allowed to take the food. Normal animals choose the right cup even after several minutes have passed. Those which have suffered bilateral ablation of the prefrontal areas choose at random, even when the cups have been hidden for the shortest time possible. The integrity of areas *9b* and 10 is necessary for the correct performance of this type of test. The interference of other stimuli, especially visual, plays an important part in this loss of memory, as the animals are able to perform this kind of test correctly if they are kept in the dark during the interval of delay. Monkeys and rats can solve correctly any problem that does not involve symbolic concepts, provided that all the factors are present at the same time. If, on the contrary, the solution requires remembrance of a recent experience, the animal fails completely, as its mental processes lack continuity.

[1] RUCH, T. A., and SHENKIN, H. A., *J. Neurophysiol.*, **6**, 349, 1943.

Emotional changes are conspicuous after frontal lobectomy. Animals with an irritable temperament become placid and good-natured. The tip of the frontal lobe and the frontal projection of the dorsal median nucleus of the thalamus apparently play an important part in the integration of emotional reactions. It has been suggested, however, that section of the uncinate fasciculus connecting the frontal and temporal lobes is of more importance in the causation of emotional changes after frontal lobectomy than the interruption of frontothalamic pathways.[1] Some evidence in favor of this hypothesis is given by the results of orbital leukotomy in man (section of the white fibers and disconnection of the cortex of the orbital aspect of the frontal lobe, *i.e.*, area 13), in which extroversion, euphoria, and increased motor activity are even more marked than after the usual operation.

Lesions of the frontal lobes produce in man disturbances similar to those observed in animals. Many cases of destruction of these lobes by pathologic processes and by trauma have been observed, and since Moniz[2] performed the first frontal lobotomy in cases of melancholia and anxiety, psychosurgery[3] has progressed considerably.

In man frontal lobectomy (extirpation of the frontal lobe) causes more extensive and permanent

[1] REITMAN, F., *Am. J. Psychiat.*, **103**, 238, 1946.

[2] MONIZ, E., "Tentatives opératoires dans le traitement de certaines psychoses," Paris, 1936.

[3] "Psychosurgery consists in operation upon the anatomically intact brain for the purpose of relieving mental abnormalities." (FREEMAN, W., and J. W. WATTS, *Ann. Rev. Physiol.*, **6**, 517, 1944.)

alterations than lobotomy (leukotomy), an operation that consists in cutting the connections between the prefrontal areas and other parts of the brain, with minimum damage to the cortex. Immediately following the operation, during 2 to 5 days, there is a period of lethargy and disorientation. The patient is in a condition of placid indifference, but must be fed and cleaned as if he were a baby. Gradually periods of stereotyped hyperactivity appear; the movements have the typical aimlessness and perseveration observed in experimental animals. In the succeeding months personality develops and the patients "grow up." Any emotional stress that existed, such as anxiety, obsessions, compulsions, and hypochondria, disappears and is replaced by a sense of happiness and well-being (euphoria). Indolence, distractibility, tactlessness, and boastfulness are prominent in the initial stages but tend to diminish and disappear as time passes. "A patient who has satisfactorily recovered from a psychosis through psychosurgery, gives the impression of a pleasant and enthusiastic, if somewhat immature individual, whose willingness to fall in with the 'set' of the other person's attitude makes him an agreeable companion. People on whom this operation has been performed are generous, steady, fairly reliable workers, friendly and unembarrassed at being interviewed by strangers. They take things as they come."[1] These subjects also "live in a perpetual present"; they are very susceptible to external stimuli, so that their interests are not in themselves but in the outside world. There is a certain quantitative relationship between the severity of the symptoms produced by the operation and the amount of frontal lobe separated from the rest of the brain.[2] If the section is performed so that 85 per cent of the frontal connections are severed, the patient cannot properly organize his behavior; there is lack of initiative, of constructive imagination, and of self-control. Even when recovery is satisfactory, the subjects are immature and have a puerile personality.

The frontal lobes play an important part in the process of "long-circuiting,"[3] *i.e.*, of interposing chains of internuncial neurons in a nerve path, thus increasing the opportunities for association between specialized cortical areas. This process is of great significance in the integration of complex reactions. The circuits through the frontal lobe are not specialized. When they are suppressed, no specific sensory or motor deficit follows, such as occurs when area 3-1-2 or area 4 is destroyed, but the number of possible circuits for association is considerably reduced. To a certain extent long paths can be built up again with what is left of the cortex after frontal lobotomy. Cobb[1] attributes the loss of recent memory to the destruction of these frontal association pathways. "Perhaps remote memories become associated with so many cortical mechanisms that no one operation will obliterate them." Improvement in recent memory, as in other faculties, would be due to the rebuilding of the lost circuits in the available areas. The cortex in this respect has a great degree of "plasticity."

[1] FREEMAN, W., and J. W. WATTS, "Psychosurgery," Charles C Thomas, Springfield, Ill., 1942.

[2] FREEMAN and WATTS, "Psychosurgery."

[3] FULTON, J. F., "Muscular Contraction and the Reflex Control of Movement," Williams & Wilkins, Baltimore, 1926.

THE BASAL GANGLIA

The main basal ganglia are the caudate nucleus and the putamen (which, together with the fiber tracts they include, form the corpus striatum), and the globus pallidus; the amygdaloid nucleus and the claustrum also form part of the complex. These nuclei discharge into several brain-stem nuclei, the subthalamic nucleus of Luys, the substantia nigra, the red nucleus, and the reticular substance. The pathways are short; they connect neighboring nuclei with each other or with the cortex. The anatomic paths have not all been established, and sometimes a connection is known only through its physiologic effects.

Connections of the basal ganglia. (Fig. 483.)

Caudate nucleus. Afferent fibers: areas 4*s*, 8*s*, and 9 of the frontal lobe and 2*s* of the parietal lobe; branches from the corticospinal tract; median nuclei of the thalamus. Efferent fibers: putamen and globus pallidus.

Putamen. Afferent fibers: areas 4 and 6; branches from the corticospinal tract; median nuclei of the thalamus; caudate nucleus. Efferent fibers: globus pallidus.

Globus pallidus. Afferent fibers: frontal (6) and parietal cortex; corpus striatum; thalamus. Efferent fibers (lenticular bundle): frontal cortex (4); thalamus (n. ventralis anterior, centrum medianum); subthalamic nucleus; hypothalamus; substantia nigra; red nucleus;

[1] COBB, S., "Borderlands of Psychiatry," Harvard University Press, Cambridge, Mass., 1943.

pontobulbar reticular formation; motor nuclei of cranial nerves; globus pallidus of the opposite side (Meynert's decussation).

Subthalamic nucleus of Luys. Afferent fibers: globus pallidus. Efferent fibers: nuclei near the substantia nigra; ipsilateral and contralateral red nuclei.

Substantia nigra. Afferent fibers: areas 4 and 6; globus pallidus. Efferent fibers: red nucleus; mesencephalic motor nuclei.

Red nucleus. (*a*) Parvicellular portion (upper two-thirds). Afferent fibers: dentate nucleus; thalamus; globus pallidus; frontal lobe. Efferent fibers: ascending, lateroventral nucleus of the thalamus (connected with areas 4 and 6); descending, reticular substance. (*b*) Magnocellular portion (lower one-third). Afferent fibers: contralateral nucleus interpositus. Efferent fibers: rubrospinal and rubro-olivary tracts.

Excitability. Stimulation of the putamen and the globus pallidus in monkeys produces no motor responses. On the other hand it inhibits phasic movements and postural reflexes that should result from simultaneous stimulation of the motor cortex. The caudate nucleus must be intact to obtain this effect.

Ablation. Localized lesions of the basal ganglia are not followed by sensory or motor disturbances. Those that destroy the striatum produce transient clumsiness in movements and slight rigidity on the opposite side. Involuntary movements (choreoathetosis) have been observed in chimpanzees after destruction of the caudate nucleus.[1] Simultaneous ablation of area 6 and the striatum produces tremor and choreoathetosis in chimpanzees; also a greater degree of rigidity than if area 6 alone is removed.

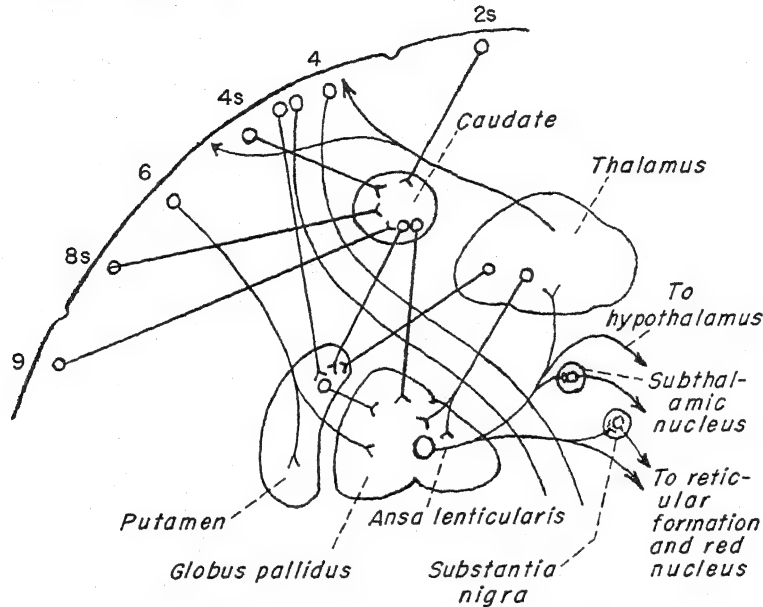



Fig. 483. Diagram of interconnections between cortex and basal ganglia.

Corticostriatal connections. Localized strychninization (Dusser de Barenne) has demonstrated the close functional relation existing between the basal ganglia and certain areas of the cortex.[2] Strychninization of area 8*s* increases electrical activity in the head of the caudate nucleus; that of area 4*s* increases it in the body; and that of area 2*s* increases it in the tail of the nucleus. Area 6 projects to the putamen and globus pallidus, and area 4 to the putamen, but neither of these areas projects to the caudate nucleus. Stimulation of the inhibitory areas is transmitted to the striatum, thence to the globus pallidus, and thence through the lateroventral nuclei of the thalamus to area 4. The interruption of any part of this circuit suppresses the inhibitory effect of stimulation of areas 8*s*, 4*s*, and 2*s* on the motor cortex (Fig. 484). Cortical

[1] Kennard, M. A., *J. Neurophysiol.*, **7**, 127, 1944.

[2] Dusser de Barenne, J. G., and W. S. McCulloch, *J. Neurophysiol.*, **3**, 311, 1941; Dusser de Barenne, J. G., H. W. Garol, and W. S. McCulloch, *Assoc. Res. Nerv. Ment. Dis.*, **21**, 246, 1942.

inhibitory areas, however, also discharge impulses into the bulboreticular formation.

Clinical observations. Much of our knowledge on the basal ganglia has been obtained by careful observation of patients suffering from lesions in these centers. The signs of such lesions can be divided into (*a*) hyperkinesia (increased motility) due to lack of inhibition of lower centers (these are therefore "release" phenomena); (*b*) hypokinesia, or absence of movement.

Hyperkinesia consists in involuntary movements and rigidity.

Involuntary movements are stereotyped but of varying amplitude. If they are ample and performed relatively slowly, they are called "athetosis"; those of lesser excursion and quicker rhythm are called "chorea." Athetosis has been observed in cases with lesions of the striatum, the globus pallidus, and the lateroventral nuclei of the thalamus, *i.e.*, any of the links in the inhibitory corticostriatal circuit. Localized lesions of the subthalamic nucleus of Luys produce athetosis with very ample movements of the ipsilateral upper limb (hemiballismus).[1] Ablation of the cortical centers (areas 4 and 6) relieves these involuntary movements but causes a certain degree of paresis.[2]

Tremor is an outstanding symptom in Parkinson's disease, a condition in which lesions have been found in the globus pallidus, the substantia nigra, and the neighboring areas. Tremor can involve the head or the limbs; in the hand it provokes typical movements similar to those made when counting coins or rolling pills. The tremor ceases when the patient performs voluntary movements and is therefore called a resting tremor, but it is not really so, because it also disappears during sleep. Parkinsonian tremor is a postural tremor, just as cerebellar tremor is an adjunct of voluntary or phasic activity. Removal of the motor cortex also suppresses tremor; the latter is therefore due to "release" of the cortical centers from inhibitory influences.

Tremor with the characteristics of Parkinsonian tremor has been observed in monkeys after the production of lesions in the tegmentum of the upper brain stem. Tremor was abolished by extirpation of areas 4 and 6, but reappeared when motility was recovered. Interruption of corticopallidal connections did not modify the tremor.[1] A similar effect is caused by injury to the subthalamus.[2]

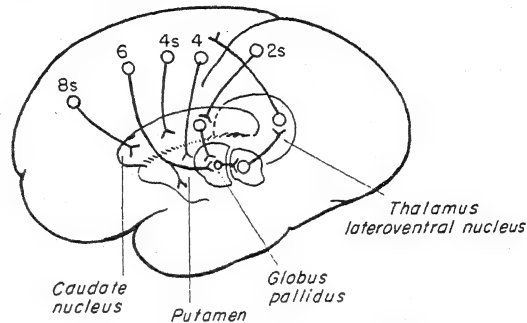



FIG. 484. Diagram of corticostriate inhibitory circuits. (*After Dusser de Barenne.*)

Tremor is an outstanding disturbance after Forel's fields have been damaged. Apparently these are the crossroads where cerebellar, cortical, and subcortical paths that are of importance for the control of involuntary movements meet.

Rigidity can be extremely marked in Parkinson's disease; it involves all muscles but is not accompanied by hyperreflexia. Section of the dorsal spinal roots suppresses rigidity in the deafferented segments. This fact would seem to demonstrate a myotatic-reflex component in rigidity.

Hypokinesia is most evident in the absence of "associated movements." Normal persons perform a series of involuntary (associated) movements when they perform a voluntary movement, *e.g.*, the swinging of the arms when walking and the movements of the face and hands when talking. These movements are reduced or may be completely absent in patients with Parkinson's disease. They walk with their arms hanging stiffly and their faces void of expression; they are "poker-faced." Another feature of hypokinesia is the slowness with which movements are initiated. These symptoms are due, at least in part, to rigidity. Hypokinesia has been provoked in cats by lesions in the caudal part of the hypothalamus

[1] The patients perform movements which are similar to those of an athlete throwing a weight, hence the term used.

[2] BUCY, P. C., "The Precentral Cortex," University of Illinois Press, Urbana, Ill., 1944.

[1] WARD, A. A., W. S. MCCULLOCH, and H. W. MAGOUN, *J. Neurophysiol.*, **11**, 317, 1948.

[2] WHITTIER, R., and F. A. METTLER, *J. Comp. Neurol.*, **90**, 319, 1941.

and the neighboring tegmentum of the midbrain. The animals were somnolent and reluctant to move; they remained in any posture that was passively given them.[1] Their condition was similar to that of catalepsy. Lesions in the caudal hypothalamus and subthalamus in monkeys diminish activity and cause inertia of such a degree it is necessary to feed and care for the animals as if they were helpless.[2] Injury to the mesencephalic tegmentum has the same effect.[3] In man lesions in the rostral brain stem produced by encephalitis are followed by lethargy and by lack of spontaneous movements and of emotional expression[4] (see Chap. 88).

Summary. The extrapyramidal corticostriatomesencephalic motor system integrates involuntary (automatic) movements and posture. This integration takes place on three levels: (*a*) cortical; (*b*) striatal; (*c*) mesencephalic. Spasticity and rigidity are observed when integration at any of these levels is suppressed, but it is greater when the cortex and striatum are simultaneously destroyed and greatest when all control of the higher centers has been removed by a section below the red nucleus (decerebrate rigidity). Besides this inhibition of tonus, movements initiated in the motor cortex are also inhibited; ablation of the premotor cortex results in perseveration or persistence of movements after they have accomplished their end and in aimless hyperactivity. Interruption of the corticostriatal inhibitory circuit results in choreoathetosis, and mesencephalic lesions provoke tremor. These symptoms are due to "release" of the cortex from the inhibitory component contributed by this system. The system is to a considerable degree bilateral in function, and no specific focal areas have as yet been located.

[1] Ingram, W. R., R. W. Barris, and S. W. Ranson, *Arch. Neurol. Psych.*, **35**, 1175, 1936.

[2] Ranson, S. W., *Arch. Neurol. Psych.*, **41**, 1, 1939.

[3] Peterson, E. W., *et al.*, *J. Neurophysiol.*, **12**, 371, 1949.

[4] Von Economo, C., "Encephalitis Lethargica. Its Sequelae and Treatment," Oxford, New York, 1931.

BIBLIOGRAPHY

Association for Research in Nervous and Mental Disease, *Proc.*, vol. 21, "The Diseases of the Basal Ganglia," 1942.

Bard, P., *et al.*, Patterns of Organization of the Nervous System, *Research Publ., A. Nerv. & Ment. Dis.*, **30**, 1952.

Bucy, P. C., "The Precentral Motor Cortex," University of Illinois Press, Urbana, Ill., 1944.

Cobb, S., "Borderlands of Psychiatry," Harvard University Press, Cambridge, Mass., 1943.

Fulton, J. F., "Frontal Lobotomy and Affective Behavior," Norton, New York, 1951.

Fulton, J. F., "Physiology of the Nervous System," 3d ed., Oxford, New York, 1949.

Fulton, J. F., Somatic Functions of the Central Nervous System, *Ann. Rev. Physiol.*, **15**, 305, 1953.

Penfield, W., and T. Rasmussen, "The Cerebral Cortex of Man," Macmillan, New York, 1950.

CHAPTER 83

The Cerebellum

The cerebellum receives fibers from the vestibular division of the eighth nerve, the vestibular nuclei, and the spinocerebellar tracts. The oldest connections of the cerebellum phylogenetically are the vestibular and spinal, but with the development of the somatic musculature and the cerebral hemispheres, the cerebellum has developed and established new connections. In mammals, especially in primates, the greater number of the afferent fibers of the cerebellum come from the pontine nuclei and, through these, from the cerebral cortex (corticopontocerebellar path). The cerebellum sends impulses to the midbrain, the spinal cord, and the cerebral cortex; every cortical region which projects to the cerebellum receives a cerebellar projection. The cerebellum is, therefore, a station in the proprioceptive path, part of the postural reflex mechanism, and an important component of the mechanism for cortically induced movements.

Recent advances in the embryology and comparative anatomy of the cerebellum, registration of potentials in centers and nerve paths evoked by stimulating the cerebellar cortex and nuclei, and registration of cerebellar action potentials following excitation of afferent paths and nerve centers, have been of considerable importance in furthering the functional analysis of this part of the brain. The work of Larsell and that of Dow are of outstanding value in this respect. Before considering the functions of the cerebellum, it will be necessary to describe briefly its structure.

Structural organization. In the lamprey *Petromyzon*—a primitive fish—the cerebellum is a small bilateral organ developed from the vestibular nuclei, with a few spinal, medullary, and tectal connections. In amphibians, the body of the cerebellum develops in a median and anterior position with regard to the vestibular cerebellum, which is formed by the two lateral lobes.

In reptiles the lateral lobes (auricular lobes) are joined together, and in birds they form the flocculonodular lobe. The fissura posterolateralis separates this lobe from the body of the cerebellum. This is the first fold to appear in the course of embryonic development, and it separates two parts of different ontogenetic and phylogenetic antiquity. In the higher vertebrates this fissure, therefore, establishes a primary division of the cerebellum into two parts: (*a*) the flocculonodular lobe; (*b*) the body of the cerebellum. In reptiles at a later stage of development a second fold, the fissura prima, divides the body of the cerebellum into the anterior and posterior lobes. A third fold, the fissura prepyramidalis, divides the posterior lobe into two parts: the anterior portion is formed by the ansiform and paramedian lobes; the posterior portion is formed by the pyramis, uvula, and paraflocculus.

In the cerebellum of mammals, three parts can be distinguished which differ in their ontogenetic and phylogenetic significance, their connections with other nerve centers, and their functional importance (Fig. 485): (*a*) the archicerebellum; (*b*) the paleocerebellum; (*c*) the neocerebellum.

The archicerebellum consists of the flocculonodular lobe. It receives impulses from the labyrinth through the vestibular division of the eighth nerve and the vestibular nuclei in the medulla. Its efferent fibers end in the vestibular nuclei (Fig. 486).

The paleocerebellum (a term in which the vestibular cerebellum is sometimes included) is formed by the anterior lobe (lingula, lobulus centralis, and culmen) and the posterior part

of the posterior lobe (pyramis, uvula, and paraflocculus). Its afferent fibers come from the spinal cord (dorsal and ventral spinocerebellar tracts), the cuneate and gracile nuclei, and the nuclei of the fifth, ninth, and tenth nerves. The lingula and uvula also receive fibers from the vestibular nuclei. The anterior lobe, moreover, receives impulses arising in tactile receptors and the somesthetic cortical centers. Efferent cortical fibers are axons of the Purkinje cells; they end in the cerebellar nuclei. Those from the uvula and lingula end in the nuclei of the roof of the fourth ventricle (n. fastigii); efferent fibers from these nuclei form the fasciculus uncinatus, which crosses the mid-line and ends in the contralateral vestibular nuclei and the reticular formation of the medulla. Other efferent cortical fibers of the paleocerebellum end in the nucleus interpositus (emboliform and globose nuclei of the higher primates), situated laterally to the nuclei fastigii. The efferent fibers of the nucleus interpositus end in the large-cell portion of the red nucleus, whence the rubrospinal tract conveys cerebellar impulses to the motor cells in the anterior horn of the spinal cord. The anterior lobe also projects to the parietal cortex.

The neocerebellum, phylogenetically the last part of the cerebellum to appear, is considerably developed in primates. It comprises the lobulus simplex, declive, lobulus ansiformis, tuber, and lobulus paramedianus, *i.e.*, the anterior part of the posterior lobe between the fissura prima and the fissura prepyramidalis. The afferent fibers come from the motor and premotor cortex (areas 4 and 6 in the frontal lobe) and apparently also from the temporal cortex. These fibers end in the nuclei of the pons. Fibers emerging from these nuclei cross the mid-line and enter the cerebellum in the brachium pontis, ending in the cerebellar cortex, mainly in the hemispheres. These fibers form the corticopontocerebellar path, and they are the most numerous contingent of cerebellar afferent fibers in primates. The neocerebellar cortex also receives impulses from the auditory and visual receptors and their cortical centers. The axons of the Purkinje cells—the only efferent cells here, as in the rest of the cerebellar cortex—end in the

Fig. 485. Diagram of the cerebellum of primates showing the functional divisions of the cerebellum, *i.e.*, vestibular spinal, and neocerebellum. (*Dow, R. S., J. Neurophysiol., vol. 5, p. 121, 1942.*)

dentate nucleus, situated laterally to the nucleus interpositus. The efferent fibers of this nucleus go to the ventrolateral nucleus of the thalamus of the opposite side, and thence cerebellar impulses are carried to areas 4 and 6 in the either directly (from the cortex of the flocculus) or indirectly through the roof nuclei (from the cortex of nodulus, lingula, and uvula). This reflex arc is important in the maintenance of body equilibrium and posture.

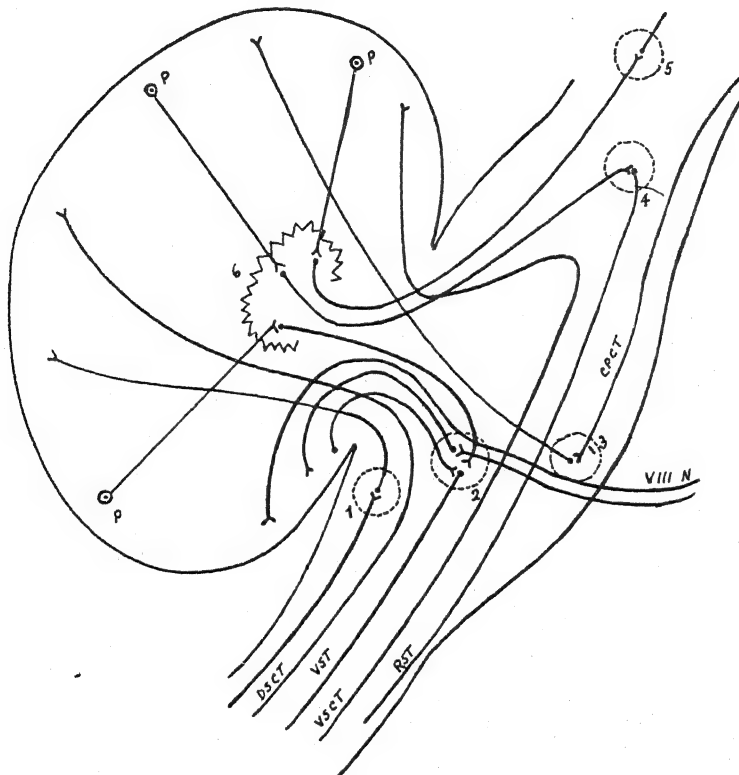



FIG. 486. Afferent and efferent fibers of the cerebellum. 1, cuneate and gracile nuclei; 2, vestibular nuclei; 3, pontine nuclei; 4, red nucleus; 5, ventrolateral nucleus of the thalamus; 6, dentate nucleus; *P*, Purkinje cells; *DSCT*, dorsal spinocerebellar tract (Flechsig's); *VSCT*, ventral spinocerebellar tract (Gower's); *VST*, vestibulospinal tract; *RST*, rubrospinal tract; VIII *N*, vestibular division of the eighth nerve; *CPCT*, corticopontocerebellar tract.

frontal cortex, and to the visual and auditory cortex. Many fibers end in the small-cell portion of the red nucleus; perhaps these cells project to the thalamus and cortex.

From a functional point of view, also, the cerebellum can be considered as consisting of three parts, which coincide approximately with the anatomic parts already described, but with some overlapping: (*a*) the vestibular cerebellum; (*b*) the spinal cerebellum; (*c*) the neocerebellum.

The vestibular cerebellum (flocculonodular lobe, lingula, and uvula) receives impulses from the labyrinth, either directly from the vestibular division of the eighth nerve or through the vestibular nuclei. It sends fibers to these nuclei and the reticular formation of the opposite side

The spinal cerebellum (anterior lobe and posterior part of the posterior lobe) receives impulses from proprioceptive receptors through the spinocerebellar tracts, and the dorsal spinal tracts through the cuneate and gracile nuclei. The efferent impulses are relayed through the nucleus interpositus and the red nucleus, and finally end on the spinal motor cells. These impulses are of importance in the control of tonus and posture.

The neocerebellum (anterior part of the posterior lobe) is the terminal station of the corticopontocerebellar pathway. It sends impulses through the dentate nucleus and the ventrolateral nucleus of the thalamus to the motor and premotor cortex. These impulses

are necessary for the control of voluntary movements.

There is no strict separation between the three parts; on the contrary, there is a certain degree of overlapping. Thus the lingula and uvula have vestibular and spinal connections, and the anterior lobe is connected with frontal and parietal cortex and the spinal cord. Overlapping also takes place at the level of the cerebellar nuclei. Thus the roof nuclei receive and emit fibers of vestibular and spinal significance, and the median part of the dentate nucleus near the nucleus interpositus projects to the thalamus and to the red nucleus, *i.e.*, to the cortex and the spinal cord.

The vestibular and spinal parts of the cerebellum form Sherrington's "head ganglion of the proprioceptive system." The neocerebellum is part of the mechanism of cortically induced movements.

EXTIRPATION OF THE CEREBELLUM

Early in the last century, Rolando, and later Flourens, described the effects produced by extirpation of the cerebellum: tremor, ataxia, and disturbances in speech. At the end of the century Luciani distinguished three stages following the operation:

1. At first, extensor tonus is increased (rigidity), and there are periodic fits of opisthotonus (contraction of spinal muscles) and convulsions. Sherrington noted that the disturbances of decerebrate rigidity are similar to those following extirpation of the cerebellum, and that this operation, or section of the cerebellar peduncles, increases their severity. He attributed these disturbances to suppression of cerebellar inhibitory impulses, *i.e.*, they are "release" phenomena of the same nature as decerebrate rigidity.
2. A few days later, the following signs appear: (*a*) *asthenia*, *i.e.*, muscular weakness; (*b*) *atonia*, or rather *hypotonia*, *i.e.*, low muscle tonus due to depression of the myotatic reflex (in cats and dogs this does not occur unless the vestibular nuclei are also damaged; in monkeys and the higher primates, cerebellar ablation provokes considerable loss of tonus); (*c*) *astasia*, *i.e.*, faulty coordination of movements (asynergia). There is no orderly combination of excitation and inhibition of synergic and antagonistic muscles, and there are many signs of disturbed motility: *dysmetria*, or errors committed in the range of voluntary movements (the subject overreaches an object or does not attain it); *decomposition of movements*, which are halting and not performed in orderly and balanced sequence; *tremor*, involuntary oscillations beginning when voluntary movements are started (intention or kinetic tremor) and increasing as the movement progresses; *adiadochokinesia* (Babinski), *i.e.*, difficulty in performing rapidly alternate opposite movements. All these disturbances are seen in walking. The patient walks as if he were drunk (*cerebellar ataxia*). Speech is also disturbed; words are alternately dragged out or emitted explosively and split up into fragments. Faulty coordination of the resonating apparatus gives a nasal intonation to the voice.
3. As time passes the disturbances diminish as a result of gradual compensation, but they do not disappear completely.

Unilateral extirpation of the cerebellum is followed by the same disturbances as bilateral removal, but limited to the operated side. Unilateral increase in the tonus of the extensor muscles causes the trunk to be bent toward the operated side. The head is twisted, with the occiput and the eyes deviated toward the operated side. The subject appears to be looking at the lesion.

FUNCTIONAL LOCALIZATIONS IN THE CEREBELLUM

The cerebellar cortex has a uniform structure; areas with different cytoarchitecture, such as there are in the cerebral cortex, have not been described in the cerebellum. At first the predominant idea was that the cerebellum acts as a whole and that no discrete part of the cerebellum corresponds to a definite part of the body. There is now definite evidence of functional and somatotopic localization in the cerebellum, obtained from the results of (*a*) localized ablation; (*b*) stimulation of the cerebellar cortex or nuclei; and (*c*) potentials evoked in the cerebellum by stimulating cortical or other nerve centers or paths, and in cortical centers by stimulating the cerebellum. The flocculonodular lobe forms part of the vestibular apparatus, the anterior lobe takes part in the regulation of tonus and posture, and the neocerebellum integrates voluntary

movements. There is also a fairly clear-cut topographic localization of sensory and motor representations in the cerebellar cortex. There is overlapping of neighboring areas, both in functional and somatotopic localization.

Flocculonodular lobe. The anatomic connections of this lobe with the vestibular division of the eighth nerve and the vestibular nuclei indicate the existence of a functional relation with the vestibular apparatus. Moreover, stimulation of the eighth nerve provokes the appearance of large action potentials in the flocculonodular lobe, uvula, lingula, and n. fastigii, *i.e.*, in the whole vestibular cerebellum.

Experimental destruction of the nodulus and flocculus in primates[1] causes disturbances in equilibrium. The animal has difficulty in standing and adopts a broad base; its trunk and head sway from side to side, and when walking it progresses with hesitating movements and easily falls down. The syndrome is sometimes called "trunk ataxia," but this is an incorrect term, as the reflexes in the limbs, which contribute to maintain the equilibrium of the body, as well as the reflexes in the trunk, are disturbed. There is no tremor or alteration in postural reflexes, and no asynergia in voluntary movements. The destruction of the uvula alone produces slight and transitory disturbances in equilibrium, but if it is destroyed together with the flocculonodular lobe, the severity of the disturbances is increased. Ablation of the lingula, and of the part of the anterior lobe that has vestibular connections, is followed by exaggerated labyrinthine reactions and disturbances in equilibrium, which do not occur if the labyrinths have been destroyed first; there are no changes in the postural reflexes of the limbs. Removal of the pyramis, uvula, and nodulus suppresses motion sickness in dogs, without disturbing reflex vomiting produced by deteriorated food or apomorphine.[2] Symptoms produced by the extirpation of the flocculonodular lobe decrease as time passes; compensation is established fairly rapidly. Carrea[3] confirmed these observations and found that ablation of the paraflocculus increased the severity of the disturbances caused by extirpation of the vestibular cerebellum (flocculus, nodulus, lingula, and uvula), but had no effect if it was the only part removed. Extirpation of the rest of the vermis produced "trunk ataxia."

[1] Dow, R. S., *Arch. Neurol. & Psychiat.*, **40**, 500, 1938.
[2] Bard, P., C. N. Woolsey, R. S. Snider, V. B. Mountcastle, and R. B. Bromiley, *Federation Proc.*, **6**, 72, 1947.
[3] Carrea, R. M. E., *Federation Proc.*, **5**, 15, 1946.

In man medulloblastomas located in the nodulus, which later invade all the vestibular cerebellum, produce the same conditions as are observed after experimental extirpation of the flocculonodular lobe.[1] The patients, who are usually children, cannot maintain their equilibrium. When lying down they move their limbs normally, but they have "trunk ataxia," *i.e.*, they stand swaying on a broad base and progress with the hesitating walk of a drunkard. There is no tremor, and there are no changes in postural tonus.

The flocculonodular syndrome is due to absence of cerebellar inhibition of the labyrinthine reflexes. There is hypermetria in vestibular reactions, this being the cause of the loss of equilibrium (Dow).

Anterior lobe. According to Connor,[2] extirpation of the whole anterior lobe is followed by serious disturbances in postural reflexes. Muscular tonus and tendon reflexes are increased. The shortening and lengthening reactions, the positive supporting reaction, and the reflexes that maintain and restore posture are exaggerated; opisthotonus is also observed. Tonic labyrinthine and neck reflexes provoke remarkable changes in posture. If the operation is performed in decerebrate animals, rigidity increases considerably. As time passes, the disturbances decrease in intensity, but a coarse intention tremor appears in the neck and forelimbs. There is also "vasomotor ataxia," *i.e.*, vascular reactions to changes in temperature are exaggerated. Other visceral reflexes, such as those that evacuate the bladder and the rectum, and the pilomotor reactions are also increased. These effects show that cerebellar inhibition acts not only on somatic reflexes but also on visceral ones. Stimulation of the anterior lobe inhibits postural tonus in the homolateral limbs. The effect is remarkable in the decerebrate preparation, in which rigidity is suppressed. After stimulation has ceased, there is considerable "rebound" and extensor tonus is greatly increased.

[1] Bailey, P., and H. Cushing, *Arch. Neurol. & Psychiat.*, **14**, 192, 1925.
[2] Connor, G. J., *Proc. Soc. Exper. Biol. & Med.*, **47**, 205, 1941.

Posterior lobe. Extirpation of the posterior lobe is followed by only a few transitory disturbances in the cat and dog. In primates, the severity of the condition increases as encephalization is greater; thus it is more severe in baboons than in monkeys, and even more so in the chimpanzee. There is muscular weakness, decreased resistance to passive movements (hypotonia), disturbances in walking, and difficulty in performing movements learned by training. Bilateral extirpation produces more marked effects than unilateral, in which disturbances are observed only on the damaged side.

Somatotopic localizations. There are two sensory-effector areas in the cerebellum. One is situated rostrally in the anterior lobe and the lobus simplex; the other caudally in the paramedian lobulus, vermis, and pyramis. These areas receive impulses from the periphery, as can be shown by stimulating the receptors and recording potentials evoked in the cerebellar cortex. Thus, in cats and monkeys stimulation of tactile receptors revealed ipsilateral representation of the cutaneous surface in the anterior lobe,[1] and in monkeys also in the paramedian lobulus.

The method of evoked potentials has also served to demonstrate the close interrelation between the cerebral cortex and the cerebellum. Discrete areas of the cerebellar cortex project to, and receive impulses from, discrete areas in the cerebral cortex. Thus, in the monkey stimulation of the face, arm, and leg areas of the motor cortex evokes potentials in the face, arm, and leg areas in the contralateral lobulus simplex and anterior lobe.[1] The second somatic area, located laterally to the main sensory area in the parietal cortex and rostrally to the auditory area, is projected bilaterally on the caudal sensory-effector area of the cerebellum.[2] Stimulation of the ansiform lobe increases the spontaneous electrical activity of the contralateral motor area. The general rule is that a cortical center which projects to a definite area in the cerebellum receives impulses from that area.

Stimulation of the sensory-effector areas of the cerebellar cortex evokes localized movements and has suppressor and facilitating effects on cortically induced movements, and on postural and spinal reflexes.

There is definite, though overlapping, somatotopic localization. Sensory, cortical, and effector representation coexist in the same area. In the rostral area, the tail, hind limb, forelimb, neck, and face are represented respectively in the

FIG. 487. Diagram of the primate cerebellum. *a*, principal divisions and connections; *b*, somatotopic localizations in the cerebellar cortex. (*Dow, R. S., Biol. Rev., vol. 17, p. 179, 1942.*)

[1] SNIDER, R. S., and A. STOWELL, *J. Neurophysiol.*, **7**, 381, 1944; *Anat. Rec.*, **94**, 489, 1946.

[1] ADRIAN, E. D., *Brain*, **66**, 289, 1943; *Brit. M. J.*, **2**, 137, 1944.

[2] DOW, R. S., *J. Neurophysiol.*, **5**, 121, 1944.

lingula, lobulus centralis, culmen, and lobulus simplex. In the caudal area, the face is represented in the upper, the forelimb in the middle, and the hind limb in the lower folia of the paramedian lobule, and the hind limb and tail in the pyramis. There are overlapping visual and auditory areas in the tuber extending into the lobulus simplex.

Electric stimulation in unanesthetized animals by means of previously implanted electrodes, or in decerebrate cats, dogs, and monkeys,[1] evokes immediate and delayed movements following the pattern of localization described above. Thus, tail movements are obtained by stimulation of the lingula, the lower folia of the paramedian lobule, and the pyramis. Hind-limb movements are produced by stimulation of the lobulus centralis, the lower folia of the paramedian lobe, and the pyramis. Forelimb movements are evoked from the culmen and the medial folia of the paramedian lobule, and neck, face, and masticatory movements from the lobulus simplex and the upper folia of the paramedian lobe. When stimulation is discontinued there is considerable rebound, *i.e.*, the opposite movements are performed. The effect is predominantly ipsilateral. Stimulation near the mid-line affects mainly the proximal muscles, and stimulation of more lateral sites affects mainly distal muscles.

Suppressor and facilitatory effects may be obtained by stimulation of the same area, the result depending on the frequency of stimulation. Stimulation of the anterior lobe in the cat at rates of 50 to 300 per second inhibits decerebrate rigidity and the myotatic reflex; reducing the frequency of stimulation to 5 to 10 per second increases spasticity.[2] In the lower mammals and birds inhibitory effects predominate. In primates facilitation of spinal reflexes and cortically induced movements are easily obtained, especially with low frequency of stimulation.[3]

Inhibitory impulses from the anterior lobe and the paramedian lobulus, which inhibit cortically induced movements and postural and spinal reflexes, are relayed by the fastigial nuclei to the reticular formation, and from there to the spinal cord. Inhibitory impulses from the cerebral cortex also converge to the reticular formation; efferent fibers from the reticular formation are therefore a final common path of cortical and cerebellar inhibition.

[1] Clark, S. L., *J. Neurophysiol.*, **2**, 19, 1939; Hampson, J. L., C. R. Harrison, and C. N. Woolsey, *Federation Proc.*, **4**, 31, 1945; **5**, 41, 1946.

[2] Moruzzi, G., *Boll. Soc. ital. biol. sper.*, **24**, 397 and 753, 1948; **26**, 125, 1950.

[3] Nulsen, F. E., S. P. W. Black, and C. G. Drake, *Federation Proc.*, **7**, 86, 1948; Snider, R. S., and H. W. Magoun, *J. Neurophysiol.*, **12**, 335, 1949.

Facilitation of cortically induced movements evoked by stimulation of the caudal cerebellar sensory-effector areas takes place in the cortex; the impulses are conducted by fibers in the superior cerebellar peduncle.[1] Stimulation of the fastigial nuclei activates the electroencephalogram; therefore, facilitating impulses probably ascend to the cortex not only by the superior cerebellar peduncle but also along reticular relays.[2] Facilitatory impulses from the anterior lobe are mediated by the fastigial nuclei, probably to a bulbospinal system of facilitation.[3]

After a comparative study of the cerebellums of man, monkey, cat, dog, sheep, pig, rabbit, and opossum, correlated with the physiological data available, Woolsey[4] has concluded that the following homologies exist between the different parts of the human cerebellum and those of animals: The amygdala in man is the homologue of the subhuman paraflocculus. The paramedian lobule, usually homologized with the amygdala, corresponds to the lobulus biventer and lower part of the inferior semilunar lobule of man. By analogy with cat and monkey, the somatotopic localization within these paramedian homologues would be as follows: leg in the lower division, and arm in the upper division, of the lobulus biventer; face in the lower part of the inferior semilunar lobule. The latter is a part of the lobus medius. The rest of the lobulus medius is composed of the rostral part of the inferior semilunar lobule (crus II), the superior semilunar lobule (crus I), and the posterior part (lobulus simplex) of the quadrangular lobule.

Electrical activity of the cerebellar cortex.[5] The spontaneous electrical activity of the cerebellar cortex, recorded with relatively coarse electrodes, shows waves of extraordinarily high frequency, about 150

[1] Rossi, G., *Arch. fisiol.*, **10**, 389, 1912; Walker, A. E., *J. Neurophysiol.*, **1**, 16, 1938.

[2] Moruzzi, G., and H. W. Magoun, *Electroencephalog. & Clin. Neurophysiol.*, **1**, 455, 1949.

[3] Snider and Magoun, *loc. cit.*

[4] Woolsey, C. N., *Ann. Rev. Physiol.*, **9**, 525, 1947.

[5] Moruzzi, G., J. M. Brookhart, and R. S. Snider, *Federation Proc.*, **8**, 113, 1949; Johnson, H. C., K. M. Browne, J. W. Markham, and A. E. Walker, *Proc. Soc. Exper. Biol. & Med.*, **73**, 97, 1950; Brookhart, J. M., G. Moruzzi, and R. S. Snider, *Federation Proc.*, **9**, 18, 1950; *J. Neurophysiol.*, **13**, 465, 1951; **14**, 181, 1951.

per second. There are also spike discharges of approximately the same frequency. Both spikes and waves are originated in the Purkinje- and granule-cell layers. The spike discharges are more sensitive to ischemia than the waves. Local strychninization increases the frequency of intrinsic units to 400 to 500 per second, but it produces synchronization with low frequency (10 to 30 per second) of the waves.

Summary. The vestibular cerebellum (flocculonodular lobe and parts of the paleocerebellum) integrates labyrinthine reflexes essential for the maintenance of the equilibrium of the body. The paleocerebellum (anterior lobe and parts of the posterior lobe) integrates postural reflexes and regulates muscle tonus. The neocerebellum, highly developed in primates, is in close reciprocal relation with the cerebral cortex for the integration of movements.

BIBLIOGRAPHY

Dow, R. S., The Evolution and Anatomy of the Cerebellum, *Biol. Rev.*, **17**, 179, 1942.

Fulton, J. F., "Functional Localization in the Frontal Lobes and Cerebellum," Oxford, New York, 1949.

Larsell, O., The Cerebellum. A Review and Interpretation, *Arch. Neurol. & Psychiat.*, **38**, 580, 1937.

Luciani, L., "Il cervelleto. Nuovi studi di fisiologia normale e patologica," Florence, 1891.

Luciani, L., "Fisiologia dell'uomo," 6th ed. (edited by S. Baglioni), vol. 3, Società Editrice Libraria, Milan, 1923.

Moruzzi, G., "Problems in Cerebellar Physiology," Charles C Thomas, Springfield, Ill., 1950.

Sherrington, C. S., "The Integrative Action of the Nervous System," Yale University Press, New Haven, 1926.

Woolsey, C. N., Somatic Functions of the Nervous System, *Ann. Rev. Physiol.*, **9**, 525, 1947.

CHAPTER 84

Visceral Motor Innervation. The Autonomic Nervous System

THE NERVOUS SYSTEM controls and coordinates visceral functions in two ways; it exercises a continuous or tonic action, and it discharges impulses that have stimulatory or inhibitory effects in physiologic emergencies. The structural and functional architecture of visceral sensibility does not differ fundamentally from that of somatic sensibility (see Chap. 73). On the other hand, visceral motor innervation differs in several respects from somatic motor innervation. Thus the visceral motor neurons are not situated in the central nervous system, but peripherally in ganglia which in some cases are found close to or within the effector organ. Moreover visceral effectors are not so dependent on nerve impulses as somatic effectors; they continue to function after denervation,[1] although their response is slower, more prolonged, and less well adjusted than when they are innervated.[2]

Historical note. Galen[3] described seven cranial nerves. The sixth nerve, which sent fibers to the larynx and the viscera, comprised the glossopharyngeal, vagus, and accessory nerves. It also included the cervical, thoracic, and abdominal sympathetic, with three ganglia: one situated under the cranium, formed by the superior cervical of the sympathetic and that of the vagus; a second in the upper part of the thorax (the stellate ganglion); and a third in the abdomen (celiac ganglion). Galen also saw the thoracic rami communicantes.

[1] Piloerector muscles, salivary and sweat glands, the adrenal medulla, and the neurohypophysis are exceptions to this rule.

[2] Hormones play an important part in the control of visceral functions.

[3] GALEN, "De usu partium corporis humanis."

In the sixteenth century, Etienne (1545) and Eustachius (1552) differentiated the vagus nerve from the sympathetic. In the following century, Willis (1621–1675) revised the anatomy of the cranial nerves and distinguished 10, the vagus nerve being the eighth. He gave the name of "intercostal nerve" to the sympathetic and believed that it had its main origin in the cranium but also received spinal roots at each intercostal space. This nerve, together with the vagus nerve, innervated the viscera; visceral motility, however, was not under voluntary control, but directed by the "cerebellum," a term that included the pons and medulla. The thoracic origin of the "intercostal nerve" was demonstrated by Pourfour du Petit (1727), who cut the nerve in the neck and saw the effect this operation produced in the eye (later known as the Bernard-Horner syndrome). He concluded that "animal spirits" must therefore travel in the nerve from the thorax to the head and not in the opposite direction, as was generally believed. Winslow gave the name of "great sympathetic" to the "intercostal nerve," because through the viscera, it established "sympathies" or reciprocal influences between the different organs.

Different functions have been attributed to the ganglia. Willis thought they were reservoirs of "animal spirits," and according to Lancisi (1718), Morgagni believed that they acted as pumps which forced the "animal spirits" along the nerves. Johnstone (1764) maintained that they could not be pumps because they had no muscular tissue and therefore could not contract. He thought they must act as filters which retained motor impulses and sensory impressions, thus giving an explanation of the fact that the former escape voluntary control and the latter are imprecise. Meckel (1749) and Scarpa (1779)

considered that they were simply centers for the distribution of nerve fibers. Winslow, Monro (1783), and later Bichat (1801) considered the ganglia to be "small brains," with all the functions of the central nervous system. Bichat[1] classified bodily functions into two groups, those corresponding to "animal life" (now called somatic), controlled by the central nervous system, and those of "organic life" (now called visceral), controlled by the ganglionic system; these systems, although separate and independent, being connected by many fibers.

In the second half of the last century the knowledge of visceral innervation progressed considerably. Remak (1854) discovered the unmyelinated fibers and saw that the ganglia were connected with the spinal nerves by two types of fiber: (*a*) the *ramus communicans sympathicus*, gray and soft, formed by unmyelinated fibers or by fibers with only a very thin myelin sheath, which arises in the ganglion and is distributed to the periphery; (*b*) the *ramus communicans spinalis*, white and firm, which is formed by myelinated fibers and connects the spinal cord with the ganglion. At the same time (1851) Claude Bernard, in the course of his researches into the origin of animal heat, discovered the vasomotor nerves. Budge and Waller (1851) described the ciliospinal center in the upper thoracic segments of the spinal cord, and Bernard[2] showed that stimulation of the floor of the fourth ventricle provoked glycosuria and polyuria, thus demonstrating the influence of the central nervous system on visceral function.

Gaskell (1856) made a careful study of the distribution in the spinal roots and in cranial nerves of the fine myelinated fibers that form the white rami. His work, confirmed and extended by Langley and his associates, finally established the structural architecture of visceral motor innervation. More recently Karplus and Kreidl (1909–1924) discovered the centers in the diencephalon (hypothalamus) that control organic functions, and evidence has now been obtained that cortical centers also exert an influence on the activity of the viscera, thus giving factual support to the suggestion, put forth by Hughlings Jackson in 1876, that there is cortical representation of visceral function.[3]

[1] BICHAT, F. X., "Anatomie générale appliquée à la physiologie et à la médecine," Paris, 1801.

[2] BERNARD, C., "Leçons de physiologie expérimentale," Baillière et fils, Paris, 1855.

[3] LANGLEY, J. N., Sketch of the Discovery in the XVIIIth Century as Regards the Autonomic Nervous System, *J. Physiol.*, **61**, 1, 1926; SHEEHAN, D., Discovery of the Autonomic Nervous System, *Arch. Neurol. & Psychiat.*, **35**, 1081, 1936.

ORGANIZATION OF VISCERAL INNERVATION

Gaskell[1] observed that in some of the ventral roots there were not only the large myelinated motor fibers, but others, also myelinated but much finer, with a diameter varying between 1.8 and 3.6 μ and exceptionally 4 μ. They are found in all the thoracic ventral roots, in the first, second, and third lumbar, in the second and third sacral, and sometimes in the first and fourth sacral, ventral roots. They are also found in the oculomotor (third), facial (seventh), glossopharyngeal (ninth), and vagus (tenth) nerves. They are not found in other spinal roots or cranial nerves. Later work has confirmed the results of Gaskell's pioneer observations, but small differences in the distribution of the fibers are found in different species and individual cases. Thus, in man white rami occur from the first thoracic to the second lumbar, and the third sacral to the fourth sacral; in the monkey, dog, and cat the first thoracic has no white ramus in some individuals, but the third lumbar and, in some cases, the fourth contain white sympathetic fibers; the first, second, and third sacral nerves also contain fine fibers. In the rabbit white rami are found from the first thoracic nerve to the fifth lumbar and from the second to the fourth sacral.

These fibers have been called "preganglionic fibers" by Langley, because they run from the cranial nerve centers and spinal cord to the ganglia. The axons of ganglionic neurons innervate visceral effectors (smooth muscle and glands); they are known as "postganglionic fibers." They have no myelin sheath, or only a very thin myelin sheath that does not extend throughout the whole length of the fiber, except those emitted by the ciliary ganglion, which are completely myelinated. Their diameter is seldom greater than 2.5 μ (Fig. 488).

The preganglionic fibers can be divided into three groups: (*a*) the *cranial* outflow, the neurons of which are situated in the tegmentum of the midbrain, the pons, and medulla, the fibers forming part of the oculomotor, facial, glossopharyngeal, and vagus nerves (fibers going from the hypothalamus through the hypophyseal stalk to the hypophysis are also included in this group); (*b*) the *thoracolumbar* outflow, the neu-

[1] GASKELL, W. H., "The Involuntary Nervous System," Longmans, New York, 1916.

rons of which are situated in each segment of the spinal cord from the first thoracic to the second lumbar, and the fibers in the corresponding ventral roots; (*c*) the *sacral* outflow, with neurons in the second and third sacral segments and fibers in the corresponding ventral roots.

tem" to the cranial and sacral divisions, and that of "autonomic nervous system" to the whole. Visceral innervation has also been called the nervous system "of organic life" or "of the ganglia" (Bichat), "of vegetative life" (Reil), and "the involuntary nervous system" (Gaskell).

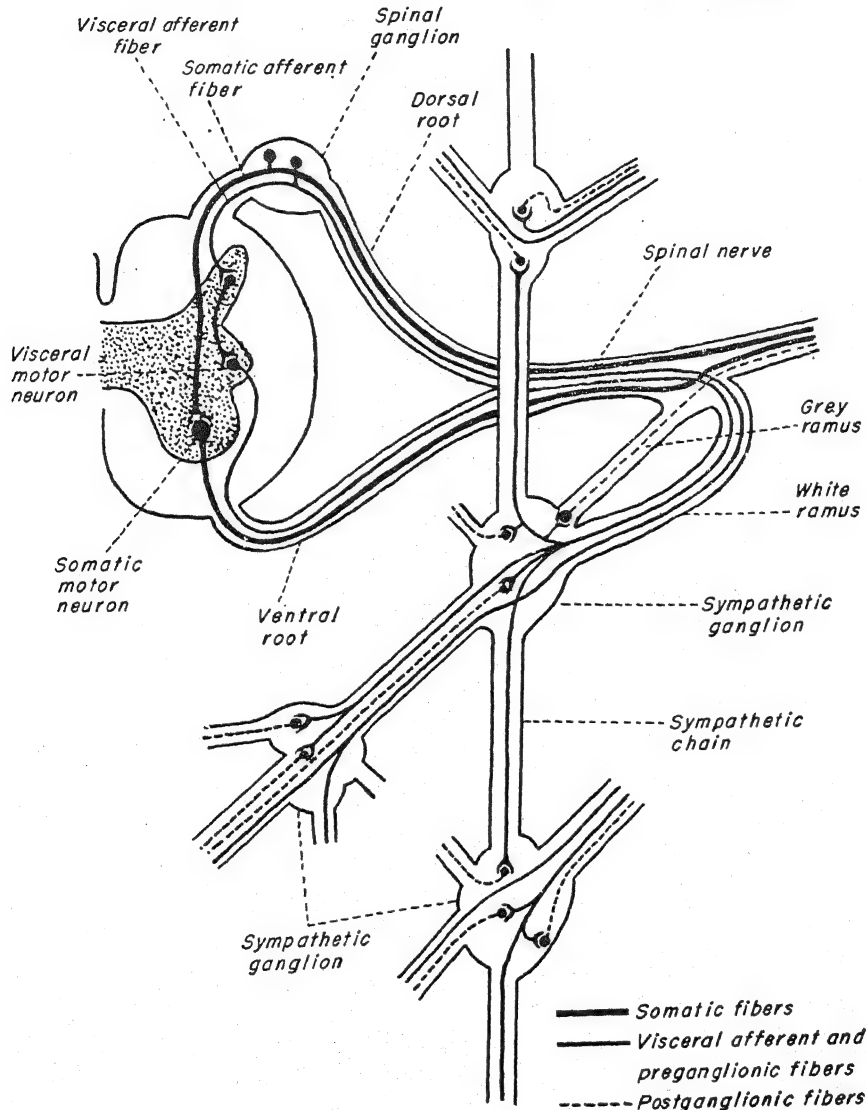



Fig. 488. The visceral reflex arc and sympathetic chain.

The first and last groups usually, but not in all cases, produce effects opposite to those produced by the second group. Langley[1] has given the name of "sympathetic system" to the thoracolumbar outflow, that of "parasympathetic sys-

[1] Langley, J. N., "The Autonomic Nervous System," Part I. Heffer. Cambridge. 1921.

At one time visceral innervation was supposed to constitute a separate and different nervous system from that which innervated somatic structures. It is now understood that there is only one nervous system.

According to Gaskell and Langley, the autonomic nervous system with its two divisions

(sympathetic and parasympathetic) is exclusively motor and peripheral. This restricted concept of visceral innervation is no longer currently accepted. The viscera are innervated

1. *Afferent neurons,* situated in the dorsal-root ganglia, or the corresponding ganglia in the facial (geniculate ganglion), glossopharyngeal (g. petrosum), and vagus (g. nodosum)

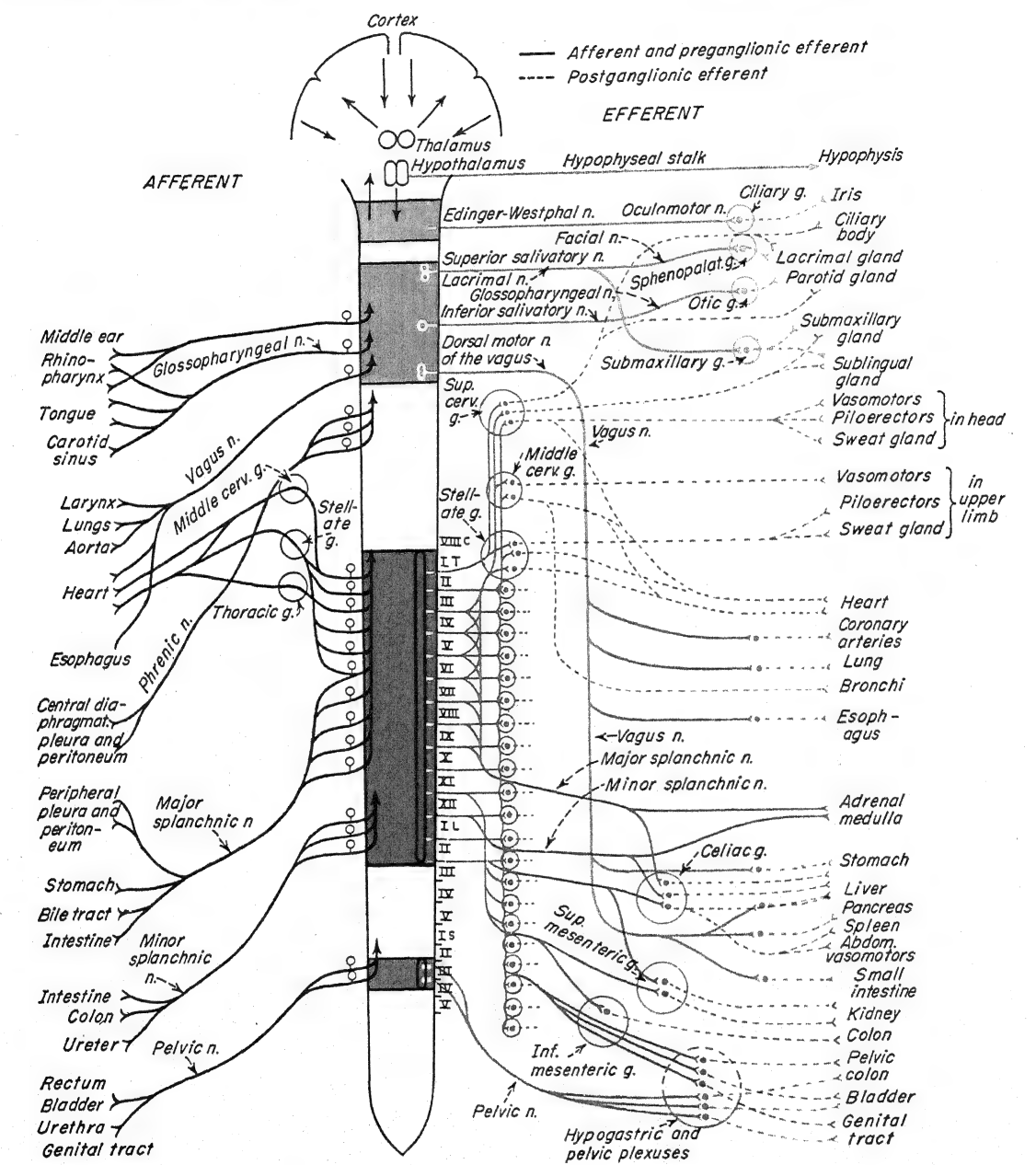



FIG. 489. Diagram of afferent and efferent fibers and centers of the visceral nervous system. *n.*, nerve; *g.*, ganglion.

in a similar way to the somatic structures, *i.e.*, by a spinal reflex arc and higher centers, with internuncial neurons connecting the different levels of integration. The elementary visceral reflex arc (Fig. 488) is made up of

nerves. The afferent fibers are usually fine and unmyelinated.

2 *Preganglionic neurons,* situated in the intermediolateral column of the spinal cord and the motor nuclei of the oculomotor nerve

(Edinger-Westphal nucleus), facial (superior salivatory nucleus), glossopharyngeal (inferior salivatory nucleus), and vagus (dorsal motor nucleus) nerves. The axons of these neurons are myelinated and form the white rami communicantes which end in the sympathetic ganglia, or the visceral efferent fibers of the cranial nerves ending in peripherally situated parasympathetic ganglia.

3. *Motor neurons* in the ganglia, which emit postganglionic fibers to the effectors.

Above the spinal level of integration, three other levels have been found: (*a*) pontobulbar integration of certain complex visceral functions, such as the blood pressure and heart rate; (*b*) diencephalic (hypothalamus) integration of still more complex patterns; (*c*) cortical integration.

Three aspects of visceral innervation can therefore be considered: (*a*) visceral sensibility (see Chap. 73); (*b*) visceral motility; (*c*) central integration of visceral activity (Fig. 489).

cervical to the second or third lumbar segment. It is formed by small cells, approximately half the size of the motor neurons in the ventral horn, having a large nucleus and relatively little cytoplasm. They are closely packed, and there is some degree of segmentation, *i.e.*, they are more numerous at a certain level of each segment. Removal of the sympathetic ganglion chain provokes chromatolysis and other signs of retrograde degeneration in these cells. Their axons are included in the ventral roots but diverge from these roots to form the white rami communicantes which go to the ganglia. Preganglionic fibers run up or down the ganglionic chain through several segments, emitting collaterals to the ganglia. They have a myelin sheath, which is sometimes lost before they end.

Sympathetic ganglia are found on each side of the spinal column throughout its whole length. They arise from the neural crest, but in the course of embryonic development they become

Afferent *Efferent*

Cortex

Thalamus

Hypothalamus → Pituitary stalk

Tegmentum → Oculomotor (III) nerve Tectal Parasympathetic

Facial (VII) nerve

Glossopharyngeal (IX) nerve →

Vagus (X) nerve

Pons

Medulla

Facial (VII) nerve

→ Glossopharyngeal (IX) nerve

Vagus (X) nerve

Bulbar Parasympathetic

Spinal cord → Thoracic I-XII

→ Lumbar I-II or III

Sympathetic

Afferent fibers in dorsal roots →

→ Sacral II or III IV or V

Sacral Parasympathetic

Visceral innervation

The Sympathetic or Thoracolumbar Division

SPINAL CENTERS. RAMI COMMUNICANTES. GANGLIA

The fibers of the thoracolumbar outflow arise in neurons situated in the intermediolateral column of Stilling, which extends from the last

separated from the dorsal-root ganglia and migrate to the periphery. Ganglionic cells are small multipolar motor neurons, which connect only with the terminals of spinal neurons. No sensory or connector cells have been found in the ganglia.

There are *three cervical ganglia* on each side of the spinal column: the superior, middle, and

inferior cervical ganglia. The middle ganglion is not found in some cases, and the inferior ganglion is usually fused with the first thoracic (stellate) ganglion. Preganglionic fibers for the superior and middle cervical ganglia leave the spinal cord in the first, second, and third thoracic ventral roots. Postganglionic fibers of the superior cervical ganglion form the plexuses of the internal and external carotid and superior thyroid artery, the superior cardiac nerve, and the gray rami for the first to the third or fourth spinal nerves. Part of the fibers emerging from the superior cervical ganglion form the internal carotid nerve or plexus which accompanies the homonymous artery. Some of these are post-synaptic unmyelinated fibers; a considerable number (one-third of the fibers leaving the SCG) are myelinated, and these have been considered preganglionic fibers ending on ganglionic cells in the carotid plexus. There are many of these cells, and they form nodes which are mostly microscopic but in some cases are visible to the naked eye.[1] Postganglionic fibers of the middle cervical ganglion form the gray ramus of the fifth cervical spinal nerve, the plexus of the inferior thyroid artery, and the middle cardiac nerve.

The *stellate ganglion* receives preganglionic fibers from the first to the eighth or ninth thoracic roots, and its postganglionic fibers form the gray rami of the sixth, seventh, and eighth cervical and first and second thoracic nerves and the plexuses of the subclavian, internal mammary, and vertebral[2] arteries; some of them also form part of the nerves in the brachial plexus.

There are ten or eleven *thoracic ganglia* as well as the stellate ganglion, which receive preganglionic fibers from the spinal nerves and send out postganglionic fibers to them. It is not possible to distinguish macroscopically the white from the gray rami because myelinated and unmyelinated fibers are included in the same fillets. Microscopic analysis of the fibers is necessary to distinguish one type from the other. Preganglionic and postganglionic fibers form plexuses surrounding the aorta, the azygos veins, and the esophagus, together with the superficial and deep cardiac and pulmonary plexuses. These plexuses receive fibers from the vagus nerve as well, and they contain ganglion cells.

Fibers from the fifth to the ninth thoracic ganglia form the *maior splanchnic nerve*, which pierces the diaphragm and ends in the celiac ganglion. Fibers from the tenth and eleventh thoracic ganglia form the *smaller splanchnic nerve*, which also ends in the celiac ganglion. In some cases a *third* or *lowermost splanchnic nerve*, ending in the renal plexus, is formed by fibers from the last thoracic ganglion. The fibers of the splanchnic nerves are myelinated preganglionic fibers, which arise in the spinal cord and not in the thoracic ganglia through which they pass.

There are usually four pairs of *lumbar ganglia* and four pairs of *sacral ganglia*. The lateral ganglionic chains frequently join each other at the caudal end in a single *coccygeal ganglion*. The preganglionic fibers of these ganglia emerge from the spinal cord in the first, second, and in some cases third, lumbar roots.

The large arteries of the abdomen are surrounded by dense plexuses formed by fibers and ganglion cells of the sympathetic and parasympathetic. The largest of these plexuses is the *aortic plexus*, which is the continuation of the thoracic aortic plexus. The celiac artery is surrounded by the *celiac*, or *solar*, *plexus*, on each side of which is a ganglionic mass, known as the celiac ganglion. This plexus receives fibers from the aortic plexus, the splanchnic nerves, the last thoracic and first lumbar white rami, and the vagus nerve. Its efferent fibers form the hepatic, gastric, pancreaticoduodenal, and splenic plexuses around the corresponding arteries, and they end in the viscera of the upper abdomen.

The superior mesenteric artery is surrounded by the *superior mesenteric plexus*, which includes ganglion cells; it is a continuation of the aortic plexus. Laterally it emits the renal and spermatic or ovarian plexuses, which surround the homonymous arteries and innervate the kidney and testes or ovaries. The *inferior mesenteric plexus*, attached to the inferior mesenteric artery, sends out fibers to the descending, sigmoid, and pelvic colon.

The lower or caudal end of the aortic plexus is joined to the *hypogastric plexus*, sometimes called "hypogastric nerve." Its caudal extremity, known as the *pelvic plexus*, also receives the pelvic nerves of the parasympathetic and

[1] Mitchell, G. A. G., *Nature, London*, **170**, 533, 1952.

[2] The plexus of the vertebral artery also receives fibers from each one of the cervical gray rami.

contains numerous ganglion cells. It innervates the pelvic viscera through the hemorrhoidal, vesical, uterine, vaginal, and prostatic plexuses.

Preganglionic and postganglionic fibers. Several methods have been used to determine the paths and endings of sympathetic fibers. The following may be mentioned:

1. Section of a ventral root is followed by degeneration of the corresponding preganglionic fibers, which can be traced by means of serial sections stained with the Marchi stain.
2. Nicotine injections provoke symptoms of generalized sympathetic stimulation, but when the drug is applied locally to a ganglion the symptoms are restricted to the territory innervated by the fibers emitted by the ganglionic neurons. After the excitatory effect has subsided, stimulation of the preganglionic fiber provokes no response, although stimulation of the postganglionic fiber provokes the usual effects. Nicotine blocks the ganglion so that preganglionic impulses are not transmitted to the ganglionic neuron. Langley used this method in tracing the distribution of the preganglionic and postganglionic fibers.
3. Action potentials of the different fibers have been registered by means of a cathode-ray oscillograph, and the excitability and speed of conduction of the fibers and the synaptic delay have been determined. The site of synaptic delay indicates where the axon of the spinal sympathetic neuron ends.

There are several types of preganglionic fibers. Four groups have been found in the cervical sympathetic with conduction speeds of 12, 8, 3, and 0.5 m. per sec., respectively. Fibers in the three fastest groups are myelinated; the slowest group is formed by unmyelinated gray fibers. Unmyelinated preganglionic fibers connect with neurons in the superior cervical ganglion that emit unmyelinated postganglionic fibers. Myelinated preganglionic fibers connect with neurons that emit myelinated or unmyelinated fibers. Approximately one-third of the fibers that leave the superior cervical ganglion are myelinated. The fastest fibers innervate the iris, the nictitating membrane, and the smooth muscle in the eyelids. Those of intermediate velocity are vasomotor and pilomotor fibers. The destination of the slow myelinated and unmyelinated fibers is as yet unknown.[1]

There are more postganglionic than preganglionic fibers. In the superior cervical ganglion the ratio is 1:16; but in other ganglia different ratios are found; *e.g.*, 1:2 in the ciliary ganglion. A preganglionic impulse can, therefore, be transmitted to several ganglionic neurons, and large action potentials are observed in the postganglionic fibers following preganglionic stimulation, owing to synchronized activation of several ganglionic neurons. Impulses sent out by a neuron in the central nervous system can thus have a widespread effect.

Stimulation of a preganglionic fiber produces effects in a large area, because it runs through several segments of the sympathetic chain and emits collaterals to the respective ganglia. On the other hand, the effects of stimulation of a postganglionic fiber are restricted to a single segment. Langley demonstrated this fact by the following experiment: the ramus communicans of a thoracic nerve was stimulated and the hair was raised along a wide area placed caudally to the stimulated root. The strength of the stimulus was then increased so as to stimulate the postganglionic fibers also, and a narrow area of piloerection appeared which was situated cephalically to the first area and separated from it by a quiescent area. Nicotinization of the sympathetic ganglion suppressed the wide caudal area of piloerection due to preganglionic stimulation, but left unchanged the narrow cephalic area due to postganglionic stimulation.

Functions of the ganglia. The structural architecture of the ganglia endows them with certain characteristics. They are formed exclusively by motor neurons; there are no sensory or internuncial neurons. All the afferent fibers come from a higher center, none from the periphery; therefore they are not reflex centers.

Kunz maintains that reflexes can be integrated in the celiac ganglion. He has observed that distention of the colon, or electrical stimulation of the mesenteric nerves, provokes a decrease in the flow of bile after all the central connections of the celiac ganglion have been severed and the duodenum cut so as to

[1] Bishop, G. H., and P. Heinbecker, *Am. J. Physiol.*, **100,** 519, 1932.

prevent the spread of impulses along the enteric plexuses.[1] Kunz and Saccomanno[2] have seen that the enteric reflex (inhibition of a proximal segment of a transected gut by dilatation of a distal segment) takes place in animals after the thoracic, lumbar, and sacral cord has been destroyed and the vagi nerves cut and after time enough has passed so that the fibers from the nerve centers have degenerated. The afferent neuron of this reflex, according to K'.nz, is situated in the wall of the gut.

The terminals of a preganglionic fiber are connected with several ganglionic neurons, and each one of these neurons is connected with several preganglionic fibers. Overlapping thus provides the structural basis for convergence, occlusion, divergence, spatial summation, and facilitation. All these phenomena have been demonstrated in the superior cervical and stellate ganglia.[3] Impulses traveling in preganglionic fibers are transmitted to the postganglionic fibers, but an antidromic volley of the latter spreads only as far as the ganglion and is not transmitted to the preganglionic fibers. Unidirectional conduction, according to Sherrington's law of forward direction, also obtains in sympathetic ganglia as in other synaptic systems.

A single impulse in the preganglionic fibers is followed by a single impulse in the postganglionic fibers; therefore the ganglionic neurons do not discharge repetitively. Repeated stimulation of preganglionic fibers at frequencies below 50 per second does not provoke repetitive discharge, but if higher frequencies are applied repetitive discharge may be observed. The frequency of physiologic impulses from the spinal sympathetic neurons never rises above 20 per second; therefore repetitive discharge of ganglionic neurons does not play a significant part in the normal transmission of impulses through the ganglia.

Repetitive stimulation, however, is followed by progressively increasing action potentials in the postganglionic fibers, because the central excitatory state spreads in the ganglion and the ganglionic neurons are recruited from the subliminal fringe as summation of the successive impulses takes place.

[1] Kunz, A., and C. van Buskirk, *Proc. Soc. Exper. Biol. & Med.*, **46**, 519, 1941.
[2] Kunz, A., and G. Saccomanno, *J. Neurophysiol.*, **7**, 163, 1944.
[3] Bronk, D. W., *J. Neurophysiol.*, **2**, 380, 1939.

The excitability of the ganglionic neuron is modified by conduction of an impulse transmitted by a preganglionic fiber, or by an antidromic volley, in the same way as conduction alters the excitability of neurons in the central nervous system. The absolute refractory period lasts approximately 2 msec.; it is followed by a phase of hyperexcitability which lasts 50 to 100 msec., coinciding with the negative afterpotential. Finally there is a period of hypoexcitability which sometimes lasts for more than 1 min., coinciding with a positive after-potential of low voltage. Repeated antidromic volleys may cause an accumulation of the phases of hypoexcitability of the ganglionic neuron, and its capacity to transmit preganglionic impulses is thus considerably depressed.

Table 115 summarizes the localization of the spinal center, the roots by which preganglionic fibers leave the spinal cord, the ganglia where they end, and the effects produced by stimulation of the different parts of the sympathetic. The facts have been obtained by observing the results of stimulation or extirpation of the different fibers in experimental animals and in human cases. The statements made should not be given too precise significance, because there are differences between species and between individuals of the same species.

The Parasympathetic or Craniosacral Division

Mesencephalic, Pontobulbar, and Spinal Centers. Parasympathetic Nerves

The preganglionic neurons of the cranial parasympathetic form a discontinuous column of nuclei called the "general visceral efferent column."[1] This extends from the tegmentum of the midbrain through the pons to the medulla and is continued as Stilling's intermediolateral column in the spinal cord. The nuclei are formed by small, closely packed neurons similar to those in the spinal centers of the sympathetic. A caudal prolongation of Stilling's column in the sacral cord contains the cells that give rise to the sacral parasympathetic (pelvic nerve).

In the tegmentum of the midbrain, close to the nucleus of the oculomotor nerve, there is

[1] Ranson, S. W., "The Anatomy of the Nervous System," Saunders, Philadelphia, 1942.

a group of small cells, the more rostral part of which fuses with the rostral end of the contralateral nucleus, forming a single median structure known as the Edinger-Westphal nucleus. The axons emitted from it form part of the oculomotor nerve and end in the ciliary ganglion. Postganglionic fibers are fully myelinated in their whole length. These fibers innervate the constrictor muscle of the pupil and the ciliary muscle. Probably the fibers for the sphincter of the pupil arise in the rostral median part of the nucleus, and those for the ciliary muscle in the caudal lateral part. The pupillary constrictor fibers apparently do not pass through the ciliary ganglion, as removal of this ganglion does not prevent the constriction of the pupil that accompanies voluntary convergence of the eyes. The fibers, possibly, end on the ganglion cells of the intraocular ciliary plexus.

The *nucleus salivatorius superior* is not well defined, nor has it been located exactly. Apparently it is formed by cells scattered in the reticular formation at the junction of the pons and the medulla. The efferent fibers of this nucleus are incorporated into the facial nerve; they separate at the level of the genu of the nerve forming the chorda tympani and end in the *submaxillary ganglion*. The postganglionic fibers innervate the submaxillary and sublingual glands and, when stimulated, provoke abundant salivary secretion, vasodilatation, and constriction of the glandular ducts. Stimulation of the nerves of these glands after partial denervation produces histological signs of activity in

Table 115. Distribution and Effects of Stimulation of the Sympathetic Division

Spinal center	*Preganglionic fibers*	*Ganglion*	*Postganglionic fibers*	*Effect of stimulation*
C_{VIII}–Th_{III}	Th_{I}–Th_{III}	Superior cervical	Plexus of internal carotid artery	Dilatation of pupil. Retraction of eyelids and nictitating membrane. Exophthalmos (not in man). Intracranial vasoconstriction
C_{VIII}–Th_{III}	Th_{I}–Th_{III}	Superior cervical	Plexus of external carotid artery	Vasoconstriction and dilatation, piloerection and sweating of head. Salivary secretion
C_{VIII}–Th_{III}	Th_{I}–Th_{III}	Superior cervical	Gray rami of C_{I-III} or C_{IV}	Vasoconstriction and dilatation, piloerection and sweating of upper part of neck
C_{VIII}–Th_{III}	Th_{I}–Th_{III}	Median and inf. cervical	Gray rami of $C_{IV-VIII}$	Same of lower part of neck
Th_{I}–Th_{III}	Th_{I}–Th_{VIII}	Stellate	Brachial plexus	Same of upper limb
C_{VIII}–Th_{I-VII}	Th_{I}–Th_{V}	Sup., med. and inf. cervical, stellate	Cardiac nerves and plexus	Increase in heart rate. Coronary dilatation
Th_{I}–Th_{II}	Th_{I}–Th_{III}	Stellate	Pulmonary plexus	Tracheobronchial dilatation. Vasomotion of lung
Th_{IV}–Th_{IX}	Th_{IV}–Th_{IX}	Thoracic	Gray rami of thoracic nerves	Vasomotion, piloerection, and sweating of trunk above the umbilicus
Th_{IX}–L_{II}	Th_{IX}–L_{II}	Thoracic and lumbar	Lumbar plexus	Same below the umbilicus
Th_{IX}–L_{II-III}	Th_{IX}–L_{III}	Lumbar	Lumbar and sacral plexuses	Same of lower limb
Th_{IV}–Th_{XII}	Splanchnic nerves	Celiac and upper mesenteric	Aortic, gastric, hepatic, splenic, upper and lower mesenteric plexuses	Abdominal vasomotion. Inhibition of gastrointestinal motility. Gastrointestinal secretion. Glycogenolysis. Adrenaline secretion (preganglionic)
Th_{XII}–L_{I}	Minor splanchnic nerve	Renal	Renal plexus	Vasomotion of kidney
$L_{I-II-III}$	L_{I-II} or III	Inf. mesent., cells in hypogastric plexus	Hypogastric plexus	Inhibition of descending sigmoid and pelvic colon. Inhibition of evacuation of rectum and bladder. Ejaculation. Uterine motility?

some of the cells, while neighboring cells remain quiescent. The existence of neurosecretory units, analogous to the motor units in skeletal muscle, has been postulated.[1]

Close to this nucleus there is another group of cells, the fibers of which also form part of the facial nerve but run in the greater superficial petrosal nerve to the *sphenopalatine ganglion*. The postganglionic fibers innervate the lacrimal glands and the small buccal and nasal glands.

The caudal portion of the nucleus salivatorius, at the level of the upper part of the nucleus ambiguus, is known as the *nucleus salivatorius inferior;* its efferent fibers form part of the glossopharyngeal nerve and run in the lesser superficial petrosal nerve to the *otic ganglion*. The postganglionic fibers (auriculotemporal nerve) innervate the parotid gland and produce the same effects in this gland as the chorda tympani produces on the submaxillary and sublingual glands.

[1] Hillarp, N. A., *Acta physiol. Scandinav.*, 17, 120, 1949.

The most important nucleus of the cranial parasympathetic is the *dorsal motor nucleus of the vagus*, in which the efferent fibers of the vagus nerve arise. These fibers end in the cardiac ganglia, the ganglia in the hilum of the lungs, the plexuses of the esophagus, the upper abdomen, and Auerbach's and Meissner's plexuses in the wall of the gut.[1] Stimulation of the vagus inhibits the heart, provokes constriction of the small coronary vessels and tracheobronchial constriction, increases gastrointestinal peristalsis and tonus, diminishes the tonus of the ileocolic sphincter, and stimulates the secretion of gastric and intestinal glands and the endocrine and exocrine secretions of the pancreas.

Fibers from the hypothalamic nuclei, which pass through the hypophyseal stalk to the different parts of the hypophysis, are now considered part of the cranial parasympathetic.

[1] Langley has suggested the name of "enteric system" for the intestinal plexuses, because it is not possible to affirm that they are part of the parasympathetic.

Table 116. Distribution and Effects of Stimulation of the Parasympathetic Division

Nerve center	*Preganglionic fibers*	*Ganglion*	*Effect of stimulation*
Hypothalamus	Hypophyseal stalk	....	Secretion of antidiuretic hormone and gonadotrophin?
Edinger-Westphal	Oculomotor (III) nerve	Ciliary Ciliary plexus?	Accommodation in near vision. Constriction of pupil
Upper salivary	Facial (VII) nerve (chorda tympani)	Submaxillary	Secretion, vasodilatation, and constriction of ducts in submaxillary and sublingual salivary glands
Lacrimal	Facial (VII) nerve (great superf. petrosal)	Sphenopalatine	Same in lacrimal, nasal, and buccal glands
Lower salivary	Glossopharyngeal (IX) nerve (lesser sup. petrosal)	Otic	Same in parotid glands
Dorsal motor	Vagus nerve	Cardiac	Decrease in heart rate; coronary constriction
Nucleus of the vagus (X) nerve	Vagus nerve	In hilum of lung	Tracheobronchial constriction
Nucleus of the vagus (X) nerve	Vagus nerve	Plexus of the esophagus	Motility of esophagus (smooth muscle)
Nucleus of the vagus (X) nerve	Vagus nerve	Semilunar, celiac, upper mesenteric, Auerbach and Meissner's plexuses	Gastrointestinal secretion. Internal and external secretion of pancreas. Gastrointestinal motility. Inhibition of ileocolic sphincter. Constriction of pancreatic and biliary ducts
SI or II–SIII or IV	Pelvic nerve (n. erigens)	Pelvic plexus	Evacuation of rectum and bladder. Pelvic vasodilatation. Erection. Secretion of glands of genital tract (prostate, Cowper's, etc.). Uterine motility?

Fibers that innervate muscles formed from the branchial clefts arise from a column of cells situated in the ventrolateral part of the reticular formation of the pons and medulla. This discontinuous mass of cells is formed by the motor nuclei of the trigeminal, and facial nerves and the *nucleus ambiguus*, which gives rise to the motor fibers of the glossopharyngeal, vagus, and bulbar portion of the accessory spinal. The fibers emerging from these nuclei innervate muscles that take part in digestive functions (sucking, chewing, and swallowing) and in the expression of emotion (facial and laryngeal muscles). The nuclei are sometimes considered as part of the visceral system, and the column they form is called the "special visceral efferent column." Their constituent cells, however, are large motor neurons with axons of great diameter and high speed of conduction, *i.e.*, with the characteristics of somatic motor neurons, and innervate striated muscles. Other authors call this group of nuclei the "lateral somatic efferent column."

The *sacral parasympathetic* consists of fibers that emerge from the cord in the second and third (in some cases also in the first and fourth) sacral ventral roots and form the *pelvic nerve*. These fibers end on the ganglion cells in the pelvic plexus which emits postganglionic fibers to the pelvic organs. Stimulation of the pelvic nerve relaxes the anal sphincter and increases peristaltic contraction of the lower part of the colon; it inhibits the sphincter vesicae and stimulates contraction of the bladder muscles exclusive of the trigonum. Parasympathetic impulses therefore play an important part in defecation and micturition. They also cause pelvic vasodilatation and erection; hence the name of "nervus erigens," which is also given to the pelvic nerve. In the female, fibers of this nerve end in the uterus, but their action is unknown, and normal labor has been observed in animals after the sacral cord has been destroyed. The spinal centers of defecation, micturition, and erection are found in the sacral segments, but it has not been possible to locate them more precisely.

Parasympathetic ganglia are placed near or even within the organ innervated by their postganglionic fibers, which are short and distributed within a limited area. The effect of parasympathetic stimulation is therefore more restricted than that of sympathetic stimulation. Moreover, a humoral mechanism (hydrolysis of acetylcholine by cholinesterase) also confines parasympathetic effects to a small area (see "Humoral factors in the transmission of visceral nerve impulses," page 1070). The distribution of parasympathetic centers and fibers and the effects of parasympathetic stimulation are summarized in Table 116.

Reciprocal Action of Sympathetic and Parasympathetic Nerves

Most of the viscera are provided with double innervation, *i.e.*, they receive sympathetic and parasympathetic fibers. There are, however, certain exceptions to this rule; thus the nictitating membrane, piloerector muscles, adrenal medulla, and sweat glands are innervated only by sympathetic fibers, which provoke contraction or secretion in these structures.

The effects produced by sympathetic impulses in viscera with double innervation are usually, but not always, opposite to those produced by parasympathetic impulses. Thus, parasympathetic stimuli provoke constriction of the pupil, and sympathetic stimulation is followed by dilatation of the pupil;[1] the heart is stimulated by sympathetic impulses and inhibited by the parasympathetic; the bronchi are dilated by the sympathetic and constricted by the parasympathetic. Many other examples of a similar nature could be given. In these cases the two systems act in a manner similar to the reciprocal innervation of antagonistic somatic muscles; the final result is due to inhibition and stimulation. The main difference, in this respect, between visceral and somatic innervation is the site of inhibition. Somatic structures receive only one type of nerve, which is exclusively excitatory, and inhibition is a central process that takes place within the central nervous system. Visceral structures receive two types of nerve, excitatory and inhibitory, and inhibition takes place in the periphery, *i.e.*, in the innervated organ. For example, flexion of a limb is provoked by impulses that travel along the nerves (excitation) of the flexor muscles (protagonists of the movement), and simultaneously the impulses in the nerves of the extensors (the

[1] Reflex dilatation of the pupil is due mainly to inhibition of parasympathetic impulses (Ury, B., and E. Gellhorn, *J. Neurophysiol.*, 2, 286, 1939). According to Kunz and Richins (*J. Neurophysiol.*, 9, 1, 1946) maximal dilatation is not obtained solely by inhibition of the circular muscle; probably it requires active contraction of the radial muscle innervated by the sympathetic. Kunz and Richins also report evidence of adrenergic (pupillodilatory) fibers arising in the ciliary ganglion.

antagonists of the movement) are diminished or suppressed (inhibition). In the visceral functions, the process is different. In the case of the heartbeat, for example, it is as follows: Tachycardia is provoked by an increase in the frequency of sympathetic (excitatory) impulses and a decrease in the frequency of vagal (inhibitory) impulses; both nerves act on the pacemaker of the heart. Bradycardia is produced by an increase in the vagal impulses and a decrease of the sympathetic impulses to the pacemaker. The importance of inhibition is demonstrated in the following experiment: The superior cervical ganglion is removed, and the pupil contracts owing to the predominance of parasympathetic impulses; if the light falling on the eye is then diminished, the pupil dilates owing to inhibition of these impulses.

In some cases both systems have similar effects; *e.g.*, stimulation of the parasympathetic provokes salivary secretion, which is also obtained by stimulation of the sympathetic. In other cases the effects of stimulation are modified by the conditions of the effector. Thus, parasympathetic stimulation usually increases gastrointestinal peristalsis and decreases the tonus of the sphincters, and sympathetic stimulation has the opposite effects; but when the gut is abnormally active, parasympathetic stimulation may inhibit the intestine, and a strongly contracted sphincter may be made to relax by stimulation of the sympathetic.

Tonic Effects of Visceral Nerves

Visceral nerves exert a continuous, or tonic, effect on the structure they innervate. If the nerves corresponding to one of the divisions of the autonomic system are cut, signs of stimulation of the other division appear. Thus, section of the cervical sympathetic or extirpation of the superior cervical ganglion, or destruction of the spinal cord at the level of the eighth cervical segment to the second thoracic is followed by the following effects due to the predominance of the parasympathetic (Bernard-Horner syndrome): (*a*) constriction of the pupil due to paralysis of the dilator of the iris and predominance of the constrictor; (*b*) narrowing of the palpebral fissure owing to ptosis of the upper lid caused by the loss of tonus in its smooth muscle; (*c*) expansion of the nictitating membrane provoked by the predominance of the fibers of the external rectus inserted on it; (*d*) the eye appears to be sunken in the orbit (enophthalmos) (*e*) vasodilatation and reddening of the skin, the conjunctiva, and the nasal mucosa, owing to paralysis of the vasoconstrictors; (*f*) dryness of the skin, due to the lack of sweat secretion; (*g*) abolishment of the piloerector reflexes in an area extending from the eye to the ear; (*h*) in some cases, a purulent secretion on the conjunctiva and nasal mucosa. In regard to (*d*) it may be noted that accurate measurements made in man have shown that the sympathetic has no influence on the position of the eye in the orbit. Stimulation of the cervical sympathetic does not cause the eye to protrude (exophthalmos), and section of the cervical sympathetic does not cause it to sink into the orbit. In certain species that have a well-developed Müller's muscle, the sympathetic has a slight effect on the position of the eyeball.[1]

The tonic effect of visceral innervation is easily demonstrated by cutting the nerves of the heart. Extirpation of the stellate ganglion is followed by bradycardia due to predominance of vagal impulses, which are no longer balanced by the sympathetic impulses. If the vagi are then cut, the heart rate increases because the continuous or tonic vagal impulses are suppressed.

Impulses travel along visceral nerves more or less continuously or periodically, a fact that has been demonstrated by registration of their action potentials.[2] These impulses are originated in different ways. Some arise in visceral receptors; thus at each heartbeat, systolic dilatation of the aorta and carotid sinuses stimulates stretch receptors which send impulses along the aortic nerve and Hering's nerve. These impulses are relayed in the nucleus of the tractus solitarius to the cardiac center, and inhibitory impulses are sent down the vagus which decrease the heart rate; at the same time sympathetic accelerator impulses are inhibited.

The activity of one center in some cases irradiates to neighboring visceral centers and causes them to discharge periodically; *e.g.*, the vasomotor center in the medulla is stimulated by the respiratory center, and at each respiratory movement volleys of impulses are sent down the vasomotor nerves. Automatic activity of the centers may also contribute to the tonic effect of visceral innervation.

[1] Pochin, E. E., *Clin. Sc.*, **4**, 79, 1939.
[2] Adrian, E. D., D. W. Bronk, and G. Phillips, *J. Physiol.*, **74**, 115, 1932.

Automatism of Visceral Nerve Centers

Visceral nerve centers are relatively independent of the centers situated at higher levels. Thus, destruction of the hypothalamus or section of hypothalamic efferent tracts is followed by ocular signs of suppression of the sympathetic (Bernard-Horner syndrome). After a few days or weeks, however, these signs disappear, and destruction of the ciliospinal center will again provoke the syndrome. After a time the effects again diminish, but extirpation of the superior cervical ganglion will then provoke a marked and lasting Bernard-Horner syndrome. Another example of recovery of tone after destruction of sympathetic centers is observed when the spinal cord is transected at progressively lower levels. After each section the blood pressure falls and then rises, but not to its original level, as the tonus in the arterial wall muscles is reduced and then partially restored.

Immediately after the destruction of a center, the absence of its impulses is made evident by the appearance of certain signs, but later the lower centers restore the normal tonus, at least in part. Visceral nerve centers do not receive impulses exclusively from higher visceral centers, but also from afferent neurons and somatic centers; recovery of tonus may therefore be due to the action of these impulses on the intact nerve centers. Ganglionic neurons, however, exert a tonic influence after they have been disconnected from the centers, and as they do not receive impulses from peripheral structures, this tonic effect must have its origin in the activity of the ganglion, *i.e.*, it must be automatic.

Most visceral structures innervated by the sympathetic and parasympathetic have a considerable degree of autonomy, and their functions are not greatly disturbed after complete denervation. This contrasts with the strict dependence of somatic muscles on their innervation, not only with respect to function (contraction) but also with respect to the maintenance of their normal structure.

The Effects of Drugs on Visceral Innervation

Certain drugs, called "sympathomimetic substances," produce effects similar to those obtained by stimulation of the sympathetic. Others produce effects similar to those due to parasympathetic stimulation; they are called "parasympathomimetic substances." Drugs of another type suppress the effects of stimulation of the sympathetic or the parasympathetic; they are known as "sympatholytic" and "parasympatholytic" drugs.

Sympathomimetic drugs are derivates of adrenaline, which together with noradrenaline is secreted by the adrenal medulla (see Chap. 85). These substances and sympatholytic drugs act on the sympathetic myoneural junction, as was suggested many years ago by Dale.[1] They do not all act with the same potency at different sites; some are mainly excitatory, others mainly inhibitory, but at sufficiently high doses they all produce all the effects of sympathetic stimulation. According to their pharmacological affinity, sympathetic receptors have been classified into two groups, α and β.[2] Stimulation of the α adrenotropic receptors by nerve impulses or drugs produces mainly excitatory effects—vasoconstriction in the viscera and skin, contraction of the uterus, ureter, and nictitating membrane, dilatation of the pupil—and one inhibitory effect, *i.e.*, intestinal inhibition. The β adrenotropic receptors are mainly associated with inhibitory effects—vasodilatation in skeletal muscle and the coronary circuit, inhibition of bronchial muscles and the uterus—and one excitatory effect, *i.e.*, increase in heart rate. Hyperglycemia and the decrease of adrenal ascorbic acid are difficult to classify. The activity of adrenaline on the α receptors is slightly greater than that of noradrenaline, but the latter has very little activity on the β receptors, which are very sensitive to isoprenaline (isopropylnoradrenaline). The order of activity of adrenaline (1), noradrenaline (2), and isoprenaline (3) on α receptors is 1-2-3; on β receptors it is 3-1-2. Sympatholytic drugs, such as ergot and its alkaloids (ergotamine, dihydroergotamine, etc.), iohimbin, and piperidylmethylbenzodioxane (Fournier's 933); suppress α effects but not β effects.

Acetylcholine is the most important of the parasympathomimetic drugs, because it is the most powerful. It acts as a chemical mediator in synaptic transmission in the ganglia, in the skeletal muscle end-plate, in parasympathetic and some of the sympathetic peripheral receptors, and there is evidence that it plays a part in

[1] Dale, H. H., *J. Physiol.*, **34**, 163, 1906.

[2] Ahlquist, R. P., *Am. J. Physiol.*, **153**, 586, 1948.

the conduction of the nerve impulse and in central synaptic transmission. Choline, pilocarpine (extracted from several species of *Pilocarpus*), and muscarine (an alkaloid in the mushroom *Amanita muscaria*) are also parasympathomimetic drugs. The latter acts only at the parasympathetic nerve endings. Atropine, the alkaloid of belladonna (deadly nightshade), is the prototype of parasympatholytic drugs; it is active at peripheral junctions, and much less at the ganglia. Drugs which inhibit cholinesterase and prevent hydrolysis of acetylcholine, such as eserine, prostigmine, di-isopropylfluorophosphate, enhance cholinergic effects or may produce them by potentiating subliminal amounts of acetylcholine released at cholinergic nerve endings.

Several drugs block sympathetic and parasympathetic ganglia, but by different mechanisms:

1. Nicotine, acetylcholine, and some of the quaternary ammoniums, *e.g.*, tetramethylammonium, depolarize the ganglionic neurons, as is shown by electrical records. The ganglionic neuron is first stimulated and discharges impulses which produce peripheral effects, *e.g.*, dilatation of the pupil in the case of the SCG; depolarization persists, and if there is a sufficiently high dose of the drug, preganglionic volleys do not evoke ganglionic potentials and the ganglionic neurons do not discharge. Large doses of eserine block a single impulse without producing depolarization, but repetitive stimulation releases sufficient ACh (which is potentiated by eserine) to block by persistent depolarization.
2. Curare and curare alkaloids (*d*-tubocurarine), tetraethylammonium, and large doses of decamethonium do not depolarize the ganglion or prevent the release of ACh by preganglionic impulses, but they block these impulses by competing with ACh for the substrate (receptor substance) on which ACh acts to produce depolarization[1] (see Chaps. 67 and 69).

Humoral Factors in the Transmission of Visceral Nerve Impulses

Elliott,[2] in his classic work on the action of adrenaline, suggested that sympathetic stimulation released it in the tissues, where it could act as a mediator for the transmission of excitation from the nerve to the effector. Shortly after, Dixon[1] suggested that muscarine could act as a chemical mediator in the peripheral transmission of parasympathetic impulses, but as this substance has not been found in animal organisms, the hypothesis did not have sufficient evidence to support it and did not receive further consideration. Later, choline and acetylcholine, substances with muscarinelike effects, were found to be normal components of the tissues. Several years after this work, Loewi[2] discovered that stimulation of the vagus in the perfused heart of the frog released a substance that produced the effects of vagal stimulation when added to the perfusion fluid in an isolated heart. The unknown active principle was called the "vagus substance."

Acetylcholine. The vagus substance has been identified as acetylcholine by pharmacologic analysis of its effects. It has not been possible to isolate it by chemical methods because it is found in blood and perfusion fluids in concentrations of the order of 1×10^{-8}.

The vagus substance loses its activity very quickly in the blood or when in contact with the tissues, because blood and tissues contain an enzyme, cholinesterase, which hydrolyzes acetylcholine into acetate and choline.[3] Eserine, and other drugs that act as anticholinesterases, such as prostigmine, di-isopropylfluorophosphate (DFP),[4] etc., protect acetylcholine and increase and prolong its effects.

The protective action of eserine has been used to demonstrate the presence of acetylcholine in venous blood and perfusates of organs in different circumstances. The experimental animal is injected with eserine or another cholinesterase inactivator, or this is added to the perfusion fluid, and the venous blood or perfusate is collected in order to test it for acetylcholine on an eserinized organ or tissue. Among the biological tests most frequently used to detect acetylcholine in blood or perfusates are contraction of the dorsal muscle of the leech, inhibition of the isolated heart of the frog, inhibition of the isolated auricle of the

[1] Paton, W. D. M., and W. L. M. Perry, *J. Physiol.*, **119**, 43, 1953; Paton, W. D. M., and J. Talesnik, *J. Physiol.*, **119**, 455, 1953.
[2] Elliott, T. R., *J. Physiol.*, **32**, 401, 1905.

[1] Dixon, W. E., *Brit. M. J.*, **2**, 1807, 1906.
[2] Loewi, O., *Pflüger's Arch. f. d. ges. Physiol.*, **189**, 239, 1921.
[3] Loewy, O., and E. Navratil, *Pflüger's Archiv. f. d. ges. Physiol.*, **214**, 678, 1926.
[4] Bullock, T. H., H. Grundfest, D. Nachmansohn, M. A. Rosenberg, and K. Sterling, *J. Neurophysiol.*, **9**, 253, 1946.

rabbit, and vasodilatation and fall in blood pressure in the cat.

The following conditions must be fulfilled in order to conclude that there is acetylcholine in an organic fluid: (*a*) atropine and curare (which do not prevent release of ACh) suppress the effects of the fluid where they suppress those of ACh; (*b*) perfusate left standing with blood or tissues in alkaline medium loses its effects, which are restored by acetylation; (*c*) boiling in acid medium does not destroy the activity of the blood or perfusate; (*d*) eserine and other drugs with anticholinesterase activity prevent destruction of the active principle and potentiate its effects; (*e*) when the fluid is tested comparatively with acetylcholine on different biological effectors, the same quantitative effects are observed; other esters of choline have greater or lesser activity on the different effectors.[1]

Liberation of acetylcholine after parasympathetic stimulation has been demonstrated in the following organs by perfusion with saline or blood (eserinized) or in eserinized animals using as indicator a denervated organ, *e.g.*, the submaxillary gland or the blood pressure: (*a*) heart; (*b*) gastrointestinal tract; (*c*) submaxillary gland; (*d*) ciliary body and iris (aqueous humor); (*e*) urinary bladder; (*f*) blood vessels of the tongue; (*g*) lung.

Stimulation of sympathetic nerve fibers liberates acetylcholine in the following cases: (*a*) in venous blood from active sweat glands and in sweat[2] (this fact explains the paradoxical action of pilocarpine, a parasympathomimetic drug, on the sweat glands, which are innervated by sympathetic fibers, reported by Langley many years ago); (*b*) in venous blood after stimulation of certain vasodilator fibers;[3] (*c*) in heart blood after stimulation of the stellate ganglion followed by coronary dilatation.[4]

These fibers have been called by Dale[5] "cholinergic fibers with muscarine action" because on stimulation they provoke the release of acetylcholine, which has effects similar to those of muscarine.

Liberation of acetylcholine in the muscle end-

[1] Gaddum, J. H., "Gefässerweitende Stoffe in der Gewebe," Thieme, Leipzig, 1936.

[2] Dale, H. H., and W. Feldberg, *J. Physiol.*, **81**, 40P, 1934.

[3] Bülbring, E., and J. H. Burn, *J. Physiol.*, **83**, 483, 1935.

[4] Folkow, F., J. Frost, K. Haeger, and B. Uvnäs, *Acta physiol. Scandinav.*, **15**, 401, 1948.

[5] Dale, H. H., *J. Physiol.*, **80**, 10P, 1933.

plate and in sympathetic ganglia (cholinergic fibers with "nicotine action"), and in stimulated nerves and nerve centers, together with its significance in the transmission of excitation, has been considered in Chaps. 67, 68, and 69.

The fact that acetylcholine exerts an effect on an organ is not sufficient proof that the organ is supplied with cholinergic innervation. Thus the nictitating membrane responds to ACh[1] yet its innervation is adrenergic.

The sweat glands are sensitive to sympathomimetic drugs. Thus intradermal injection of adrenaline or noradrenaline into the forearm or palm of human subjects, in which the regional nerves have been blocked by procaine, provokes local sweating. This response is suppressed by previous local injection of sympatholytic drugs (ergot alkaloids, etc.). This fact has led to the suggestion that sweat glands have not only cholinergic but also adrenergic innervation.[2] Intradermal injection of atropine suppresses reflex thermal sweating, or sweating induced by insulin hypoglycemia in the area of skin thus treated. Sympatholytic drugs do not have this effect when injected locally; therefore it is unlikely that the sweat glands are innervated by adrenergic fibers.[3]

Sympathin. Stimulation of the sympathetic provokes liberation on the periphery of a substance that has many of the properties of adrenaline. Loewy[4] was the first to observe this; a cardioaccelerator substance appears in the fluid of the perfused heart of a toad after the cardiac sympathetic has been stimulated. Cannon and Bacq[5] demonstrated a similar fact in intact animals. Stimulation in the cat of the sympathetic fibers which innervate the tail provoked piloerection and vasoconstriction in the tail and acceleration of the denervated heart. If the tail veins were tied, piloerection and vasoconstriction were observed but the heart rate did not vary; on reestablishing the venous circulation from the tail, tachycardia was observed. The humoral cardioaccelerator factor liberated by sympathetic stimulation was called sympathin.

Sympathin is more stable than acetylcholine and can therefore act not only locally but also

[1] Rosenblueth, A., *Am. J. Physiol.*, **100**, 443, 1932.

[2] Haimovici, H., *Proc. Soc. Exper. Biol. & Med.*, **68**, 40, 1948.

[3] Chalmers, T. M., and C. A. Keele, *J. Physiol.*, **114**, 510, 1951.

[4] Loewy and Navratil, *loc. cit.*

[5] Cannon, W. B., and Z. Bacq, *Am. J. Physiol.*, **96**, 392, 1931.

on distant organs. It can be demonstrated in venous blood proceeding from an organ in which the sympathetic has been stimulated. This blood is tested on an organ or tissue that responds to sympathomimetic substances. The following biological tests are among those most frequently used: contraction of the nictitating membrane, dilatation of the pupil, acceleration of the denervated or isolated heart, constriction of the arterioles and hypertension, relaxation of the duodenum, contraction of the denervated spleen, relaxation of the retractor penis. Denervation not only suppresses the influence of the nerve centers, but also increases sensitiveness to sympathin; therefore denervated structures should be used whenever testing a fluid for this substance. Sensitiveness can also be increased by previous treatment with cocaine.

Liberation of sympathin after stimulation of the cervical sympathetic has been demonstrated in the aqueous humor and in the venous blood from the salivary gland; also after stimulation of the stellate ganglion, the abdominal sympathetic chains, the hepatic nerves, the splenic nerves, etc. Dale has called those nerve fibers which on excitation release sympathin, "adrenergic fibers."

Cannon and Rosenblueth[1] demonstrated that stimulation of sympathetic nerves released two different sympathins. They obtained evidence for this by experiments such as the following: The nictitating membrane and the uterus of a female virgin cat were denervated and their movements were recorded. Stimulation of the abdominal sympathetic chain provoked contraction of the nictitating membrane (excitatory effect) and relaxation of the uterus (inhibitory effect). Stimulation of the hepatic nerves provoked contraction of the nictitating membrane but not relaxation of the uterus (Fig. 490). They therefore supposed that the chemical mediator (M) released in the tissues, combined with a receptor substance E or I producing excitatory

FIG. 490. Sympathins E and I. Upper tracing, contraction of the nictitating membrane. Lower tracing, contraction of denervated nonpregnant uterus. *A*, stimulation of hepatic nerves, liberation of sympathin E; *B*, stimulation of right splanchnic nerve, liberation of sympathins E and I; *C*, stimulation of right splanchnic followed by injection of adrenaline (0.3 cc., 1:200,000); *D*, stimulation of duodenohepatic nerves, liberation of sympathins E and I; *E*, stimulation of duodenohepatic nerves after cutting the duodenal branches, liberation of sympathin E. (*Cannon, W. B., and A. Rosenblueth, Am. J. Physiol., vol. 104, p. 577, 1933.*)

[1] CANNON, W. B., and A. C. ROSENBLUETH, "Autonomic Neuroeffector Systems," Macmillan, New York, 1937.

sympathin E (ME) or inhibitory sympathin (MI). Thus stimulation of liver nerves provokes the release of sympathin E almost exclusively, because the liver contains only the E receptor. In the greater part of the tissues both types of receptor substances are found, also both types of sympathin; therefore, after stimulation of the stellate ganglion acceleration of the heart would be due to release of sympathin E and dilatation of the coronary arteries to release of sympathin I.

The presence in sympathetic nerves and other tissues of a substance similar to adrenaline, which is released by excitation, has been reported by several workers.[1] More recently von Euler[2] extracted from the spleen a blood-pressure-raising substance which resembled noradrenaline and which differed from adrenaline by several of its pharmacologic properties and by its behavior in color and fluorescence reactions. Later it was found in larger amounts in extracts of the thoracic and lumbar sympathetic chain[3] and in the splenic nerves of cattle, also in the blood of man and cattle.[4] Noradrenaline has now been found in the adrenal medulla and in the blood of the adrenal vein (see Chap. 85).

A mixture of adrenaline and noradrenaline is released on stimulation of the hepatic nerves and can be found in the blood of the hepatic veins.[1] Adrenaline and noradrenaline are also released in the perfused rabbit's ear following sympathetic stimulation. Stimulation of the splenic nerve causes the appearance of noradrenaline, and in some cases of smaller amounts of adrenaline, in the blood of the splenic veins.[2] There is now general agreement that sympathin, *i.e.*, the substance released by stimulation of adrenergic nerves, is a mixture in variable proportions of adrenaline and noradrenaline, as was suggested by Bacq and Fischer.[3] Metabolic products of adrenaline oxidized in the tissues, such as adrenochrome, also seem to exert physiologic effects, *e.g.*, acceleration of blood coagulation.[4]

Preganglionic fibers
Acetylcholine
Sympathetic
Parasympathetic
Postganglionic fibers
Acetylcholine
Sympathin
Cholinesterase

FIG. 491. Diagram of chemical mediators in sympathetic and parasympathetic innervation.

Histamine. Histaminelike substances are released from stimulated nerves[5] and in the periphery by antidromic stimulation of the dorsal roots, which provokes vasodilatation.[6] There may be histaminergic nerves, as well as cholinergic and adrenergic nerves.

Summary. Stimulation of sympathetic and parasympathetic preganglionic fibers liberates acetylcholine in the ganglia, and stimulation of somatic motor fibers liberates it at the neuromuscular junction (cholinergic fibers with nicotinelike action). Stimulation of postganglionic

[1] CALABRO, Q., *Riv. biol.*, **15**, 299, 1935; GADDUM, J. H., and M. A. KHAYYAL, cited by Gaddum, *op. cit.;* CANNON, W. B., and K. LISSÁK, *Am. J. Physiol.*, **125**, 745, 1939; GADDUM, J. H., and H. KWIATKOWSKI, *J. Physiol.*, **94**, 87, 1938; **96**, 385, 1939; GADDUM, J. H., C. S. HANG, and H. KWIATKOWSKI, *J. Physiol.*, **96**, 104, 1939; BÜLBRING, E., *J. Physiol.*, **103**, 55, 1944.

[2] VON EULER, U. S., *Acta physiol. Scandinav.*, **11**, 168, 1946.

[3] *Ibid.*, **12**, 73, 1946; **16**, 63, 1948.

[4] VON EULER, U. S., and C. G. SCHMITERLÖW, *Acta physiol. Scandinav.*, **13**, 1, 1947.

[1] WEST, G. B., *Nature*, **163**, 721, 1949.

[2] PEART, W. S., *J. Physiol.*, **108**, 491, 1949.

[3] BACQ, Z., and P. FISCHER, *Arch. internat. de physiol.*, **55**, 73, 1947.

[4] BACQ, Z. M., *Proc. Roy. Soc., London, s.B.*, **137**, 300, 1950.

[5] VON EULER, U. S., and A. ASTRÖM, *Acta physiol. Scandinav.*, **16**, 97, 1948.

[6] LEWIS, T., and H. M. MARVIN, *Heart*, **14**, 27, 1927.

parasympathetic fibers liberates acetylcholine in the periphery (cholinergic fibers with muscarinelike action). Stimulation of certain postganglionic sympathetic fibers, such as those which innervate the sweat glands, and some of the vasodilator fibers also releases acetylcholine. Acetylcholine is hydrolyzed immediately by the action of cholinesterase; therefore its effects cannot spread or last very long (Fig. 491).

Stimulation of most of the postganglionic sympathetic fibers causes the release of a chemical mediator (sympathin), which is a mixture in variable proportions of adrenaline and noradrenaline (adrenergic fibers).

CENTRAL COORDINATION OF VISCERAL FUNCTIONS

Integration of visceral function is carried out at different levels of the central nervous system. The spinal cord integrates simple reflex responses, while more complex patterns are integrated in centers situated in the pons and medulla, and even more complex patterns in the hypothalamus. There is a cortical level of visceral integration, and perhaps the cerebellum also plays a part in the regulation of visceral functions.

Spinal Visceral Reflexes

A first degree of integration of visceral functions is carried out in the spinal cord. Visceral reflexes are depressed in the spinal animal during the period of shock, but later certain visceral functions are integrated at the spinal level. Thus defecation, micturition (with incomplete evacuation of the bladder), certain sexual reflexes (erection), vasomotor responses, and sweating provoked by heat, cold, and nocuous stimuli that normally cause pain are among the visceral reactions that are observed after transection of the spinal cord. The response to stimulation is sluggish, but after the reflex has developed it tends to spread and persist, giving abnormally intense and widespread responses (mass reflex). Thus distention of the bladder may cause hypertension and profuse sweating (see "The spinal animal," Chap. 81).

Medullary and Pontine Centers

Several fairly complex visceral reactions are integrated at the medullary level; *e.g.*, regulation of the blood-sugar level, the heart rate, and the blood pressure. Integration of visceral and somatic responses also takes place at this level, *e.g.*, vomiting, swallowing, and coughing.

Toward the middle of the last century, Claude Bernard[1] demonstrated that a prick with a needle ("piqûre") on the floor of the fourth ventricle provoked glycosuria. Later it was shown that mechanical or electrical stimulation must be applied in the vicinity of the dorsal nucleus of the vagus to obtain this effect. Impulses are discharged from a center situated in this spot; they travel down the spinal cord and through the splanchnic nerves to the liver and adrenals. Hepatic glycogenolysis and hyperglycemia are produced by a double mechanism: (*a*) nerve impulses pass down the splanchnic and hepatic nerves and stimulate the liver cells directly; (*b*) nerve impulses also pass down the splanchnic nerves to the adrenal medulla, and adrenaline is discharged and carried by the circulation to the liver, where it provokes glycogenolysis.[2]

If the stimulus is applied slightly above this center, hypoglycemia is provoked, perhaps by stimulation of the pancreatic islets and insulin secretion.

It has been argued that "piqûre" of the floor of the fourth ventricle does not act on pontobulbar centers but stimulates the paths descending from the hypothalamus. However, the existence of reflex centers is demonstrated by the fact that certain types of reflex hyperglycemia and hypoglycemia can be provoked after the brain stem has been cut immediately below the mesencephalon, but not after the pons and medulla have been destroyed. There is, therefore, satisfactory evidence of a pontobulbar control of the blood-sugar level.

Bernard also observed that "piqûre" provoked polyuria, and later an increase in renal elimination of salt was observed. Water metabolism and the osmotic equilibrium of the body are, therefore, controlled by mechanisms integrated in pontobulbar centers.

Sweat secretion is integrated in part in pontobulbar centers. This has been demonstrated by means of the so-called "galvanic reflexes," in which the electrical resistance of the skin is determined, because variations in electrical resistance are due to changes in the amount of sweat secreted. Some of these reflexes are local

[1] Bernard, *op. cit.*

[2] Houssay, B. A., and E. A. Molinelli. *Rev. Asoc. méd. argent.*, 37, 235, 1927.

or segmental reflexes integrated in the spinal cord; others cannot be provoked in spinal animals but are observed if the pons and medulla are intact. Another reflex of this type, the so-called "psychogalvanic reflex," is integrated in the cortex.

The bulbar vasomotor center was discovered many years ago in Ludwig's laboratory by means of the following experiment: The blood pressure of an animal was recorded, and the nerve centers were transected at different levels. No significant changes in blood pressure were observed until the section was made below the medulla, in which case the blood pressure fell. Stimulation of the upper part of the floor of the fourth ventricle caused hypertension due to vasoconstriction. Stimulation of a neighboring area caused vasodilatation and a fall in blood pressure. These vasomotor centers are located in the reticular formation of the upper part of the medulla. Reflexes provoked by centripetal stimulation of the cardioaortic and carotid sinus nerves are integrated in these centers; they cannot be obtained in the spinal animal but are intact in the bulbospinal animal.

The vomiting center is located in the dorsolateral reticular formation; if this area is destroyed, dogs no longer vomit when injected with apomorphin or given copper sulfate by mouth. An emetic trigger zone has been found in the area postrema;[1] if it is destroyed, animals cease to respond to apomorphin, although their vomiting center remains intact and they vomit after ingestion of copper sulfate. The emetic trigger zone is also part of the mechanism responsible for motion sickness; animals which vomit on being swung cease to be sensitive to swinging if this zone is destroyed[2] (see "Flocculonodular lobe," Chap. 83).

The dorsal motor nucleus of the vagus is an important center for visceral reflexes; the heart rate, bronchial musculature, and gastrointestinal motility are regulated by impulses discharged from this nucleus down the vagus. Reflexes that assure the normal evacuation of the bladder are also integrated in the medulla; the spinal animal and patients who have suffered transection of the spinal cord have permanent disturbances in micturition (see Chap. 64).

[1] WANG, S. C., and H. L. BORISON, *Arch. Neurol. & Psychiat.*, **63**, 928, 1950.
[2] WANG, S. C., and H. I. CHINN, *Federation Proc.*, **11**, 400, 1952.

Reflex regulation of the diameter of the pupil, including the consensual reflex, is integrated in the mesencephalon (Edinger-Westphal nucleus).

The Hypothalamus

The importance of subcortical centers situated above the corpora quadrigemina in the regulation of visceral functions was demonstrated many years ago. Goltz's decorticate dogs had little or no disturbance in visceral functions, but transection of the central nervous system at the level of the corpora quadrigemina caused considerable loss in the control of visceral functions. The centers that take part in this control are situated in the diencephalon, one of the oldest parts of the brain considered from the phylogenetic and ontogenetic point of view.

Anatomic organization (Fig. 492). The hypothalamus forms the floor of the third ventricle. Above, it is separated from the thalamus by the sulcus hypothalamicus. Anteriorly it is continued into the paraolfactory region of the telencephalon (the preoptic and septal areas). Its rostral limit has been established conventionally by a transverse plane passing immediately in front of the anterior border of the chiasma, thus excluding the median and lateral preoptic nuclei. The caudal limit of the hypothalamus is given by a plane passing immediately behind the mammillary nuclei, although there is no definite cleavage between the hypothalamic nuclei and those of the tegmentum of the mesencephalon. Laterally the hypothalamus is continuous with the subthalamic area, *i.e.*, the subthalamic nuclei of Luys, the zona incerta, the capsula interna, and the cerebral peduncle.

The medial forebrain bundle, which runs backward from predominantly olfactory centers (olfactory bulb, olfactory tubercle, paraolfactory area) to the tegmentum, occupies the most lateral part of the hypothalamus and permeates the lateral nuclei. It is more highly developed in the lower vertebrates than in mammals. In microsmatic animals, such as the primates, it is less important. This bundle establishes connections between the olfactory centers and the hypothalamus.

The hypophysis is placed immediately below the hypothalamus, and the pars nervosa of this gland is an outgrowth of the floor of the third ventricle. Important vascular and neural connections between the hypothalamus and the hypophysis are established through the pituitary stalk.

"The hypothalamus of mammals may, for descrip-

tive convenience, be subdivided into three regions from before backwards: 1) the pars supraoptica (in relation to the optic chiasma); 2) the tuber cinereum, to which is attached the stalk of the hypophysis, and 3) the pars mammillaris."[1]

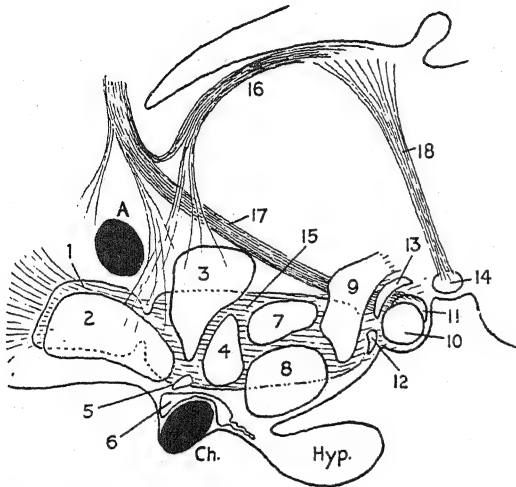



FIG. 492. Diagram of hypothalamic region (sagittal section). *A*, anterior commissure; *Ch*, optic chiasma; *Hyp*, hypophysis; 1, n. preopticus lateralis; 2, n. preopticus medialis; 3, n. paraventricularis; 4, anterior hypothalamic area; 5, n. suprachiasmaticus; 6, n. supraopticus; 7, n. hypothalamicus dorsomedialis; 8, n. hypothalamicus ventromedialis; 9, n. hypothalamicus posterior; 10, n. mammillaris medialis; 11, n. mammillaris lateralis; 12, n. premammillaris; 13, n. supramammillaris; 14, n. interpeduncularis; 15, n. hypothalamicus lateralis; 16, stria habenularis; 17, fornix; 18, fasciculus retroflexus of Meynert. The medial forebrain bundle passes through 1 and 15. (*Le Gros Clark, W. E., et al., "The Hypothalamus," Oliver & Boyd, Ltd., Edinburgh and London, 1938.*)

The following classification and nomenclature of the hypothalamic nuclei and areas, based on the work of Le Gros Clark, have been generally adopted[2] (Fig. 492).

1. Periventricular region.
   a. Periventricular system.
   b. N. preopticus periventricularis.
   c. N. arcuatus periventricularis.
2. Lateral region.
   a. Lateral hypothalamic area.
   b. Lateral nuclei of the tuber.
3. Medial rostral or supraoptic region.
   a. N. supraopticus.
   b. N. paraventricularis.
   c. N. suprachiasmaticus.
   d. N. supraopticus diffusus.
4. Medial tuberal or infundibular region.
   a. N. hypothalamicus ventromedialis.
   b. N. hypothalamicus dorsomedialis.
   c. Posterior hypothalamic area.
   d. Perifornical area.
5. Caudal medial or mammillary region.
   a. N. premammillaris.
   b. N. supramammillaris.
   c. N. mammillaris medialis.
   d. N. mammillaris lateralis.
   e. N. intercalatus.

The periventricular region has small cells distributed in rows in the wall of the third ventricle; it is continued posteriorly into the gray substance of the aqueduct. The lateral hypothalamic area has mostly small and median cells dispersed throughout it; there are also a few large cells similar to those of the supraoptic and paraventricular nuclei. These cells are more numerous in man than in other species, especially in the caudal part of the area. There are two or three small groups of cells in its ventral part, which are known as the lateral nuclei of the tuber.

Above the chiasma (the medial rostral region), there are two main nuclei: the supraoptic and paraventricular nuclei. They are formed by closely packed large cells, with vacuolae containing colloids and lipids, and the Nissl granules in a peripheral position. These nuclei have a very active circulation, and secretory activity has been attributed to them (Roussy's "neurocrinia" and Scharrer's "*Zwischenhirndrüse*"). Hypophysectomy or transection of the hypophyseal stalk causes these nuclei to atrophy owing to retrograde degeneration. Most of the cells of the supraoptic nucleus and, in man, a large proportion of the cells of the paraventricular nucleus are lost after these operations, which, however, do not cause lesions in any other hypothalamic nucleus.

There are two principal nuclei in the medial tuberal region, the n. hypothalamicus ventromedialis and n. hypothalamicus dorsomedialis, both formed by small cells. The hollowing out of this region by the infundibulum of the third ventricle is less marked in man than in other species. The posterior hypothalamic area is well developed in man; it is formed by scattered cells which give rise to the fibers of the periventricular system, which is the principal efferent pathway of the hypothalamus.

[1] LE GROS CLARK, W. E., *et al.*, "The Hypothalamus," Oliver & Boyd, Edinburgh and London, 1938.
[2] LE GROS CLARK, W. E., The Hypothalamus, *Proc. Assoc. Res. Nerv. Ment. Dis.*, **20**, New York, 1940.

The mammillary region has a median nucleus formed by large cells, and premaxillary and supramaxillary nuclei well developed in macrosmatic animals. The n. intercalatus, formed by large cells, is situated between the mammillary nuclei and the fornix, from which it receives fibers.

The afferent fibers of the hypothalamus are as follows:

1. The mammillary nuclei receive fibers from the brain stem which form the spinobulbar and pontobulbar hypothalamic tracts in lower vertebrates but which are collaterals of the median lemniscus in mammals. These fibers conduct impulses from the receptors of general sensibility. Probably some of these fibers conduct impulses from the gustatory fibers in the cranial nerves.
2. The supraoptic nucleus is connected to the vagus by an unknown path. Evidence of this is given by the fact that centripetal stimulation of the vagus provokes hypertension, which does not occur if the hypophysis is removed or the hypophyseal stalk is transected [see No. 2 (supraoptic-hypophyseal tract) under "The efferent fibers," below].
3. Fibers ascending in the periventricular system arise in the tegmentum (corpora quadrigemina). They have been well demonstrated in lower vertebrates, but less clearly in mammals. They pass through the mammillary peduncle and end in the lateral mammillary nucleus.
4. The median lemniscus sends fibers to the hypothalamus, which, together with the thalamohypothalamic fibers, conduct somatic afferent impulses. Visual and auditory impulses may reach the thalamus through the geniculate bodies and the subthalamus.
5. Fibers which come from the striatum by way of the median forebrain bundle, and from the globus pallidus (pallidohypothalamic tract) by way of the subthalamus.

The efferent fibers of the hypothalamus are as follows:

1. Fibers to the nuclei in the tegmentum, which form part of the mammillotectal tract.
2. Fibers from the supraoptic paraventricular nuclei and cells in the anterior part of the infundibular eminence, which form the supraoptic-hypophyseal tract and pass through the hypophyseal stalk into the hypophysis, mainly into the pars nervosa.
3. Fibers for lower centers, arising principally in the posterior hypothalamic area, passing down the periventricular system into the dorsal longitudinal bundle, and scattered fibers which do not form a compact bundle. By means of stimulation and transections it has been possible to establish that these fibers are situated in the tectal part of the pons and laterally in the reticular formation of the medulla; in the spinal cord they form part of the anterolateral column. They decussate at different levels and end in the visceral centers of the midbrain, pons, medulla, and spinal cord.

The hypothalamus is related to the limbic cortical system in a way similar to the relations existing between the thalamus and the neocortex.[1] Fibers in the medial forebrain bundle arise in the olfactory and paraoltactory areas (considerably developed in macrosmatic animals); fibers in the fornix connect the hippocampus with the mammillary body; fibers in the mammillothalamic tract connect the mammillary bodies with the anteroventral nucleus of the thalamus, which projects to the gyrus cinguli. These paths seem to be two-way systems, as are also hypothalamic-neocortical projections. A large unmyelinated motor projection from the posterior part of the orbital gyri (area 13) to the ventromedial and lateral nuclei of the hypothalamus has been found in monkeys and man.[2] The posterior orbital gyrus receives impulses from the afferent fibers of the vagus, relayed in the medial part of the posterior nucleus.[3]

Projections to and from the neocortex are usually relayed in intermediate nuclei, but there is a direct unmyelinated projection from the precentral neocortex. The nucleus ventralis posterolateralis projects impulses arising in splanchnic nerve fibers to ipsilateral and contralateral somatic sensory areas I and II in the region corresponding to the trunk.[4] Other impulses of splanchnic origin are relayed through the posterior hypothalamus and caudal thalamus. The hypothalamus is also connected to the frontal cortex by a two-way projection relay: (*a*) the dorsal nuclei of the thalamus; (*b*) anterodorsal nuclei of the septum (neocortical septal and septohypothalamic tracts); and (*c*) the zona incerta of the subthalamus.

The hypothalamic nuclei are interconnected by many fibers.

The hypothalamus has an active circulation supplied by arteries from the arterial circle of Willis. Popa and Fielding[5] described a portal circulation formed by capillaries in the hypothalamus, opening

[1] Le Gros Clark, W. E., *Lancet*, **2**, 353, 1948.

[2] Wall, P. D., P. Glees, and J. F. Fulton, *Brain*, **74**, 56 and 66, 1951; Beck, E., *et al.*, *Brain*, **74**, 295, 1951.

[3] Dell, P., *J. de physiol.* (*Paris*), **44**, 471, 1952.

[4] Amassian, V. E., *J. Neurophysiol.*, **14**, 433, 445, 1951; Downman, C. B. B., *J. Physiol.*, **113**, 434, 1951; **116**, 228, 1952.

[5] Popa, Q. T., and N. Fielding, *J. Anat.*, **65**, 88, 1930; **67**, 227, 1933.

into vessels in the hypophyseal stalk, which again ramify into capillaries in the hypophysis. They maintained that blood flows from the hypophysis to the hypothalamus. Later work[1] has shown that blood flows in the opposite direction, *i.e.*, from the hypothalamus to the ventral aspects of the pars distalis of the hypophysis.

Stimulation of the hypothalamus. The different parts of the hypothalamus have been stimulated by implanting electrodes and applying the stimulus when the animal has recovered from the effects of the operation and anesthetic. Occasionally, in the course of surgical operations, the hypothalamus has been stimulated in man. The optimum rate of stimulation is high compared with that for visceral effectors, and this is important because lowering the rate (without changing the situation of the electrodes) may reverse the effects provoked, *e.g.*, a pressor response may be converted into a depressor response.[2] After stimulation, signs of visceral activity or inhibition of visceral activity, mediated by the sympathetic and parasympathetic, may be observed. The response is frequently a complex pattern integrated with somatic reactions.

According to Hess and his associates,[3] it is more convenient to refer to longitudinally located zones than to specific centers for definite responses. Hess divides the hypothalamus in this respect into medial, intermediate, and lateral zones. The medial area is relatively unresponsive; the caudal hypothalamus gives mainly sympathetic responses, and the rostral areas are concerned with the integration of complex patterns coupled to somatic movements, such as defecation.

Dilatation of the pupil, contraction of the nictitating membrane, vasoconstriction, tachycardia, and other signs of sympathetic activity have been provoked by stimulation of the hypothalamus. Dilatation of the pupil in the cat is caused by two different mechanisms with separate locations; one inhibits the oculomotor (parasympathetic) center, the other stimulates the sympathetic. In the monkey the latter is the more important, because section of the cervical sympathetic diminishes or abolishes the dilator response to hypothalamic stimulation, while in the cat the inhibitory mechanism is more important.[1] Destruction of the lateral and caudal areas of the hypothalamus is followed by a Bernard-Horner syndrome, which is more marked if the destruction is bilateral but which does not last long. It would seem, therefore, that hypothalamic centers exert a tonic influence on the innervation of the pupil.

Other local effects of hypothalamic stimulation will be considered later, but widespread activation of sympathetic visceral effectors may be due to an increase in the secretion of adrenaline provoked from the hypothalamus (see Chap. 85). Release of sympathin has also been obtained by hypothalamic stimulation.[2]

Typical vagal effects on the intestine (increased peristalsis) have been observed following stimulation at the level of, or caudal to, the infundibulum; double vagotomy abolished these effects. Stimulation of more rostrally situated parts caused vasoconstriction and inhibition in the gut, followed by a marked excitatory response; since these effects were not abolished by vagotomy,[3] they were probably due to excitation and subsequent inhibition of the abdominal sympathetic. Inhibition of peristalsis in the stomach, intestine, and colon (*i.e.*, sympathetic effects) has been obtained by stimulation of the lateral areas of the hypothalamus.[4] Salivation and an increase in gastric secretion and acidity have also been reported following hypothalamic stimulation.

Stimulation of the anterior hypothalamus provokes activity in the bladder, mediated by the pelvic and hypogastric nerves. Movements of defecation have also been observed following stimulation of this part of the hypothalamus.

The effects provoked by stimulation of the hypothalamus are not due to stimulation of paths descending from the cortex; they have been obtained after decortication and degeneration of corticofugal fibers. Moreover, local cocainization of the hypothalamus abolishes them, and the administration of barbiturates

[1] Lascano-González, J. M., *Rev. Soc. argent. de biol.*, **11**, 309 and 318, 1935; Green, J. D., *Anat. Rec.*, **97**, 338, 1947; **99**, 21, 1947; **100**, 273, 1948.

[2] Hare, K., and W. A. Geoghegan, *Am. J. Physiol.*, **126**, P524, 1939.

[3] Hess, W. R., U. Brügger, and V. Bücher, *Monatschr. f. Psychiat. u. Neurol.*, **111**, 17, 1945.

[1] Weinstein, E. A., and M. B. Bender, *J. Neurophysiol.*, **4**, 44, 1941; Hodes, R., and H. W. Magoun, *Am. J. Physiol.*, **133**, 330, 1941.

[2] Magoun, H. W., S. W. Ranson, and A. Hetherington, *Am. J. Physiol.*, **119**, 615, 1937.

[3] Wang, S. C., G. Clark, and S. W. Ranson, *Am. J. Physiol.*, **130**, 81, 1940.

[4] Sheehan, D., *Am. J. Digest. Dis.* **9**, 361, 1942.

(amytal, nembutal) depresses hypothalamic excitability.

Functional patterns integrated in the hypothalamus. Several complex visceral synergies, sometimes coupled with somatic reactions, are integrated in the hypothalamus; *e.g.*, the regulation of body temperature, water metabolism, certain aspects of carbohydrate and fat metabolism and of sexual functions, general bodily activity, sleep, etc.

Temperature regulation. Over forty years ago, Isenschmidt and Schnitzler[1] suggested that the hypothalamus was concerned in heat regulation, a fact that has been given definite proof by later work. Keller and Hare[2] demonstrated that the chief central mechanism controlling heat production is located in the hypothalamus. Transection of the brain immediately caudal to the mammillary nuclei converts a homeothermic animal into a poikilothermic one. On the other hand the hypothalamic animal (*i.e.*, an animal in which the forebrain has been removed but the hypothalamus remains intact) regulates its body temperature and maintains it at a constant level. This is achieved by a complex hypothalamic mechanism that integrates visceral and somatic reactions to heat and cold. Thus an animal placed in a cold environment responds by vasoconstriction, piloerection, and increased secretion of adrenaline (visceral reactions), which diminish heat loss, and by shivering, a somatic reaction, which increases heat production. The response to heat involves activation of visceral mechanisms such as vasodilatation and sweating and the somatic response of panting (heat polypnea).

The different parts of the hypothalamus have been stimulated locally by introducing electrodes heated by means of low-voltage high-frequency currents. Warming of the anterior hypothalamus causes a fall in rectal temperature by activation of mechanisms that increase loss of heat: there is cutaneous vasodilatation and sometimes vasoconstriction in the splanchnic area. If the skin is cooled at the same time as the hypothalamus is warmed, vasodilatation is diminished or suppressed. Cooling of the anterior hypothalamus causes vasoconstriction in the skin if the cutaneous vessels are dilated, but not otherwise. Warming the anterior hypothalamus also provokes polypnea and sweating.[1] Stimulation of the lateral and caudal hypothalamus is followed by cutaneous and splanchnic vasoconstriction and dilatation of blood vessels in muscles. The latter response is mediated by cholinergic vasodilator sympathetic fibers, and it can be suppressed by atropine or sympathetic denervation of the muscles.[2] Simultaneously there is cutaneous piloerection, discharge of adrenaline from the adrenal medulla, hyperglycemia due to increased glycogenolysis, and a rise in temperature.

Bilateral destructive lesions of the rostral (preoptic) area provoke a postoperative rise in temperature in monkeys, which may reach a fatal level if it is not treated. The heat-loss mechanism is paralyzed; there is no vasodilatation or sweating, and the threshold of thermic polypnea is raised considerably. The heat-conservation mechanism is maintained and is perhaps exaggerated; heat production continues. Pentobarbital causes a fall in temperature in these animals, which is due to suppression of the activity of the heat-conservation mechanism (there is peripheral vasodilatation, and piloerection ceases) and to a decrease in heat production (shivering ceases).[3] In man lesions in the vicinity of the third ventricle produce hyperthermia due to vasoconstriction and cessation of sweating, *i.e.*, to paralysis of the mechanism controlling loss of heat.

Destructive lesions in the caudal part of the lateral hypothalamus are followed by a fall in temperature owing to damage to the centers that regulate the reactions to cold.[4] If the animals are placed in a cold environment they do not respond with shivering, cutaneous vasoconstriction, or piloerection, which increase heat production or decrease the loss of heat. These

[1] Isenschmidt, R., and W. Schnitzler, *Arch. f. exper. Path. u. Pharmakol.*, **76**, 202, 1914.

[2] Keller, A. D., and W. K. Hare, *Proc. Soc. Exper. Biol. & Med.*, **29**, 1069, 1932.

[1] Magoun, H. W., F. Harrison, J. R. Brobeck, and S. W. Ranson, *J. Neurophysiol.*, **1**, 101, 1938; Hemingway, A., T. Rasmussen, H. Wikoff, and A. T. Rasmussen, *J. Neurophysiol.*, **3**, 329, 1940; Beaton, L. E., W. A. McKinley, C. M. Berry, and S. W. Ranson, *J. Neurophysiol.*, **4**, 478, 1941; Ström, G., *Acta physiol. Scandinav.*, **20**, Suppl. 70, 1950; **21**, 271, 1951.

[2] Folkov, B., G. Ström, and B. Uvnäs, *Acta physiol. Scandinav.*, **17**, 327, 1949; Eliasson, S., B. Folkov, P. Lindgren, and B. Uvnäs, *Acta physiol. Scandinav.*, **23**, 333, 1951.

[3] Beaton, L. E., C. Leininger, W. A. McKinley, H. W. Magoun, and S. W. Ranson, *Arch. Neurol. & Psychiat.*, **49**, 518, 1943.

[4] Clark, G., H. W. Magoun, and S. W. Ranson, *J. Neurophysiol.*, **2**, 61, 1939.

lesions also destroy efferent paths from the more rostrally situated centers which control the reactions to heat, and the animals become completely poikilothermic.

After a time these disturbances are compensated to a certain extent, and the animals can maintain a fairly constant body temperature, but when they are submitted to stress, *e.g.*, when they are placed in an ice chest or in a warm room, their thermic reflexes are sluggish and the body temperature may fall or rise.

The temperature-regulating centers in the hypothalamus are brought into activity by stimulation of cutaneous receptors that are sensitive to hot and cold, and by changes in the temperature of the blood circulating through the centers. The importance of the temperature of the blood in the regulation of body temperature was demonstrated by Sherrington by means of the following experiment: The spinal cord of a dog was transected at the level of the upper thoracic segments, so as to disconnect the hindquarters of the animal from the higher nerve centers. The hind limbs were then submerged in cold water, and soon the forelimbs and head began to shiver. This response could not be of reflex origin, because all nerve paths from the hind limbs had been suppressed; therefore it must have been due to cooling of the blood as it passed through the skin submerged in cold water. Later, Barbour[1] showed that warming the carotid blood caused hypothermia and that cooling it caused hyperthermia. In the first case, warm blood arriving at the hypothalamus stimulated the centers that control the rise of body temperature and provoked excessive heat loss and a fall in temperature. In the second case, cool blood stimulated the centers that control the fall of body temperature and provoked excess heat production and reduced the loss of heat so that the body temperature was increased.

The hypothalamus seems to be of secondary importance in the control of vasomotor responses to changes in the temperature of the environment, which are mainly regulated by reflexes initiated in the superficially placed receptors sensitive to warming and cooling of the skin. Changes in the internal temperature stimulate receptors placed deeply in the skin (which provoke sweating on being warmed)[2] and

[1] BARBOUR, H. G., and A. L. PRINCE, *J. Pharmacol. & Exper. Therap.*, **6**, 1, 1914.

[2] BAZETT, H. C., *J. Applied Physiol.*, **4**, 245, 1951.

the hypothalamus. Integration of visceral (vasomotor reactions, sweating, piloerection, etc.) and somatic (shivering, panting) responses to changes in internal and also in external temperature, performed at the hypothalamic level, is of major importance in the maintenance of a constant body temperature.

Regulation of water metabolism. The maintenance of a constant body temperature is closely connected with the water balance and the osmotic equilibrium of the body. Cooling provokes concentration, and warming provokes dilution, of the blood. Destructive lesions of the hypothalamus disturb these reactions. Thus permanent polyuria is caused by destruction of the supraoptic nucleus or by transection of the supraopticohypophyseal tract or the hypophyseal stalk.[1] The syndrome of diabetes insipidus, observed in man and characterized by great thirst (polydipsia) and the daily elimination of large quantities of urine of low specific gravity, is thus reproduced experimentally. On the other hand, water diuresis can be inhibited in rabbits by stimulation of the supraoptico-hypophyseal tract[2] (see Chap. 52). This is a cholinergic mechanism. Injection of small amounts of acetylcholine into the supraoptic nuclei releases the antidiuretic hormone of the posterior pituitary; and the same effect can be obtained by intracarotideal injection of ACh in unanesthetized, atropinized dogs. Di-isopropylfluorophosphate injected into the supraoptic nuclei also produces marked and long-lasting inhibition of the rate of urine flow owing to accumulation of ACh; later (4 to 19 days) there is polyuria, probably due to an excess of ACh.[3]

Cardiovascular control. Stimulation of the anterior and middle parts of the hypothalamus as well as the posterior part may provoke an increase in blood pressure. Sometimes a fall in blood pressure is observed by lowering the frequency of stimulation without changing the position of the stimulating electrode.[4] Stimula-

[1] FISHER, C., W. R. INGRAM, and S. W. RANSON, *Arch. Neurol. & Psychiat.*, **34**, 124, 1935.

[2] HARRIS, G. W., *Proc. Roy. Soc., London, s.B.*, **22**, 385, 1947.

[3] PICKFORD, M., *J. Physiol.*, **106**, 264, 1947; DUKE, H. N., M. PICKFORD, and J. A. WATT, *J. Physiol.*, **111**, 81, 1950; DUKE, H. N., and M. PICKFORD, *J. Physiol.*, **114**, 325 and 333, 1951.

[4] BERRY, C., W. A. MCKINLAY, and R. HODES, *Am. J. Physiol.*, **135**, 338, 1942; HODES, R., and H. W. MAGOUN, *J. Comp. Neurol.*, **76**, 169, 1942.

tion of the preoptic area causes a decrease in heart rate, and stimulation of the hypothalamus causes an increase in heart rate.[1] Records of the electrical activity of the inferior cardiac nerves and the cervical sympathetic[2] show that during hypothalamic stimulation there is an increase of activity in these nerves which ceases when stimulation is stopped. Probably hypothalamic impulses are relayed in the medullary sympathetic centers. Destructive lesions of the hypothalamus cause no change in the electrical activity of sympathetic nerves, nor are they followed by changes in heart rate or blood pressure; therefore, hypothalamic centers do not exert a "tonic" effect on the circulation.

Carbohyarate metabolism. Hypothalamic mechanisms play a part in the regulation of carbohydrate metabolism through the sympathico-adrenal glycogenolytic mechanism. Certain hyperglycemic reactions, *e.g.*, that provoked by cold, do not take place or are depressed in animals after the hypothalamus has been destroyed. Lesions in the neighborhood of the paraventricular nucleus provoke hypoglycemia and hypersensitiveness to insulin (see Chap. 41).

Fat metabolism. A hypothalamic mechanism regulating fat metabolism has been postulated because in certain cases of tumors of the brain, located in the diencephalon, adiposity or emaciation have been observed. Some of the cases are characterized by adiposity together with atrophy of the reproductive organs (Fröhlich's syndrome, or dystrophia adiposogenitalis). Marked obesity has been observed following hypothalamic lesions in the dog and rat.

In the rat this condition is provoked by bilateral destruction of a zone extending backward from the ventromedial nuclei to the region dorsolateral to the mammillary bodies.[3] It is caused by interrupting connections arising in the hypothalamus and descending in the brain stem. It is not provoked by interruption of the hypothalamic paths to the hypophysis, nor is it prevented by removal of the hypophysis.

In rats[4] oxygen consumption was found to be low and RQ high. Obesity seems to be due

[1] Wang, S. C., and S. W. Ranson, *Am. J. Physiol.*, **132**, 5, 1940.

[2] Pitts, R. F., M. G. Larrabee, and D. W. Bronk, *Am. J. Physiol.*, **135**, 338, 1942.

[3] Hetherington, A. W., and S. W. Ranson, *J. Comp. Neurol.*, **76**, 475, 1942.

[4] Tepperman, J., J. R. Brobeck, and C. N. H. Long, *Am. J. Physiol.*, **133**, 468, 1941.

mainly to an enormous increase in appetite. The animals eat twice as much as the controls, but in paired feeding experiments (the experimental animal being fed the same amount as the control) the rat with the hypothalamic lesion increased in weight only slightly more than the litter-mate control. Decrease in activity is also a factor in the obesity of these animals. On the other hand, lesions in the lateral hypothalamus, at the level of the ventromedial nucleus, lead to the refusal of food, even in animals with hyperphagia caused by previous destruction of the ventromedial nucleus.[1] Ingestion of food is controlled by a dual hypothalamic mechanism; impulses integrated in the lateral hypothalamus increase hunger, which is inhibited by impulses from the more medially placed ventromedial nucleus. These reactions are integrated at the cortical level in the posterior orbital gyrus (area 13).

Sexual functions. Evidence of the importance of the hypothalamus in the control of sexual function is given not only by cases of Fröhlich's syndrome but by several other facts (see Chap. 52). Thus a bulbospinal female cat does not present the usual estral behavior whether the estrus has occurred naturally or has been provoked by the injection of estrogens; while the decorticate cat, which has an intact hypothalamus, shows no significant changes in estral behavior.[2] The nuclei that control sexual behavior have not yet been located. In species such as the rabbit, in which ovulation does not occur spontaneously, the importance of the hypothalamic-hypophyseal mechanism has been demonstrated (see Chaps. 52 and 58). Stimulation of the hypothalamus provokes liberation of the luteinizing hormone and ovulation in the rabbit; this effect is not obtained by direct stimulation of the hypophysis. This also seems to be a cholinergic mechanism, because banthine suppresses copulatory ovulation in the rabbit.[3]

Emotional expression. Integration of the visceral and somatic reactions of emotional states takes place in the hypothalamus. A cat that has been recently decorticated shows marked somatic

[1] Anand, B. K., and J. B. Brobeck, *Proc. Soc. Exper. Biol. & Med.*, **77**, 323, 1951; *Yale J. Biol. & Med.*, **24**, 123, 1951.

[2] Bard, P., *Proc. Assoc. Res. Nerv. Ment. Dis.*, **20**, 551, 1940.

[3] Sawyer, C. H., J. E. Markee, and W. H. Hollinshead, *Endocrinology*, **41**, 395, 1947; Sawyer, C. H., *et al.*, *Endocrinology*, **44**, 18 and 134, 1949; *Am. J. Physiol.*, **166**, 223, 1951.

and visceral reactions typical of intense emotion. The slightest stimulus provokes movements of flight; the animal struggles vigorously, puts out its claws, cries, and switches its tail. The hairs are erected, there is tachycardia, the pupils are dilated, the blood sugar rises, and adrenaline is secreted; *i.e.*, there are signs of intense and widespread activity of the sympathetic. This condition, which has been called "pseudoaffective" by Sherrington, and "sham rage" by Cannon, is quickly suppressed by transection of the brain immediately below the hypothalamus.[1] Normally the cortex inhibits the diencephalic centers that integrate emotional expression, and when these inhibitory stimuli are missing, emotional reactions are exaggerated. Thus the pseudoaffective state is a release phenomenon, similar to decerebrate rigidity. A chronic bulbospinal or mesencephalic animal shows only fragmentary and incomplete expression of emotion, which is in marked contrast with the violent behavior of the decorticate or hypothalamic animal.[2]

"Sham rage" has been provoked by stimulation of the lateral areas of the hypothalamus in unanesthetized cats by Ranson and his associates and has later been confirmed by many others (Hess, Masserman, etc.). Local strychninization of the hypothalamus (Dusser de Barenne) or injection of strychnine into the third ventricle[3] causes spontaneous, sometimes violent, pseudoaffective reactions and increases the response to hypothalamic stimulation. "Sham rage" and other emotional reactions provoked by stimulation of the hypothalamus do not last long. Visceral responses also usually cease soon after stimulation is discontinued. In the intact animal, on the other hand, emotional disturbances often persist for a very long time after the stimulus has ceased to act.

Extensive bilateral destruction of the hypothalamus in cats and monkeys causes stupor, alternating with pseudoaffective manifestations of rage when the animal is annoyed or restrained.[4] Localized destruction of the ventromedial nuclei in cats was followed by a marked change in the response of the animals to friendly treatment and

[1] Bard, P., *Am. J. Physiol.*, **84**, 490, 1928.
[2] Macht, M. B., and P. Bard, *Federation Proc.*, **1**, 55, 1942.
[3] Masserman, J. H., *J. Pharmakol.*, **64**, 335, 1938.
[4] Masserman, J. H., *Arch. Neurol. & Psychiat.*, **39**, 1258, 1938; Kessler, M. M., *Proc. Soc. Exper. Biol. & Med.*, **47**, 225, 1941.

handling; any interference provoked a savage and malevolent reaction.[1]

Changes in emotional behavior and other mental disturbances have been reported in cases in which the hypothalamus has been damaged. The patients presented two types of symptoms:

1. There was psychic excitation of a maniacal character, with great loquacity, absence of control of behavior with the performance of sexual misdemeanors, anxiety, fear, strong emotional reactions, and in some cases hallucinations. Intellectual deficiencies similar to those observed in cases of frontal lobotomy have been observed, although there were no cortical lesions.
2. When the area near the mammillary nuclei was found to be destroyed, the patients had been somnolent and apathetic, and had had a tendency to hypothermia.

The hypothalamus should not be considered as a mechanism for the production of emotion, but as part of the effector mechanism of emotional expression[2] which integrates visceral and somatic reactions, stimulating and perhaps inhibiting emotional activity.

General activity and sleep. The level of general activity of the cortex (sleep, somnolence, restful waking, alertness) is conditioned by impulses from the hypothalamus and other parts of the diencephalon and from the reticular formation. This will be discussed in Chap. 88, Sleep.

Stress. There is some evidence that the hypothalamus plays a part in the reaction to stress. Thus insulin, adrenaline, hypoxia, *i.e.*, stimuli which provoke stress, increase the electrical activity of the posterior hypothalamus.[3] Electrical stimulation of the posterior part of the tuber cinereum and the mammillary bodies is followed by lymphopenia similar in its course to that observed in stress.[4] Section of the pituitary stalk, however, does not prevent corticoadrenal response to acute stress.[5]

[1] Wheatley, M. D., *Arch. Neurol. & Psychiat.*, **52**, 298, 1944.
[2] Gellhorn, E., "Autonomic Regulation," Interscience Publishers, Inc., New York, 1943.
[3] Porter, R. W., *Am. J. Physiol.*, **169**, 629, 1952.
[4] De Groot, J., and G. W. Harris, *J. Physiol.*, **111**, 335, 1950.
[5] Cheng, C. P., G. Sayers, *et al.*, *Am. J. Physiol.*, **158**, 45, 1949.

Striatohypothalamic and pallidohypothalamic pathways have been found, but their physiologic significance is still unknown.

The part played by the cerebellum in the integration of visceral function is not known. It has been reported that extirpation of the cerebellum is followed by retarded gastric evacuation and severe constipation; also that stimulation of the cerebellar cortex inhibits gastrointestinal motility and exerts a depressor action on the medullary centers regulating the circulation. Several years ago Fulton[1] remarked that "there is some reason to believe that any form of representation existing in the motor regions of the cerebral cortex probably also exists in the cerebellum" and suggested that "further work may lead to a conception of ataxia and asynergia in autonomic regulation similar to that now held for cerebellar disturbances in the somatic sphere."

Cortical Integration of Visceral Functions

Visceral responses are integrated (in some cases with somatic movements) at the cortical level mainly in the rhinencephalon, but also in the neocortex.[2]

Observations made in cats, dogs, monkeys, and anthropoids and in man have shown that stimulation of several areas in the rhinencephalon evokes visceral, somatovisceral, and somatic responses.[3] These areas can be divided into two groups: (*a*) the rostral part of the anterior limbic cortex, around the genu of the corpus callosum, and (*b*) the posterior orbital gyrus, the anterior part of the insula, the anterior part of the hippocampus gyrus, and the adjacent temporal cortex. These areas are directly related with the anterior thalamus, hypothalamus, epithalamus, and pontine structures.

Stimulation of either of these areas slows down the respiratory rhythm and may produce temporary apnea. Acceleration of the respiratory rhythm is evoked from the posterior part of the anterior limbic area, from the anterior and ectosylvian gyri, and from the anterior sigmoid gyrus in cats and dogs.[4] Vasomotor responses are obtained from the same two areas that inhibit respiratory movements, independently of the respiratory effect. Pressor and depressor reactions are observed, depending on the site of stimulation and in some cases on the frequency of the stimulus; low frequencies (10 to 20 c.p.s.) evoke depressor and high frequencies (40 to 80 c.p.s.) pressor reactions. Section of both vagi abolishes or greatly reduces depressor responses but has no effect on pressor responses.

Movements associated with the ingestion and digestion of food can be evoked from these areas.[1] Well-coordinated chewing movements follow stimulation of the pyriform area and amygdala; licking, salivation, swallowing, and in some cases retching can be provoked by stimulation of the olfactory tubercle and the adjacent pyriform cortex. Inhibition of peristalsis and tonus of the pyloric antrum has been obtained by stimulating (*a*) the rostral part of the anterior limbic cortex, and (*b*) the postorbital with the adjacent insular cortex, the anterior part of the hippocampal gyrus, and the adjacent temporal lobe. Augmentation of peristalsis and tonus can also be obtained from the second of these areas.[2]

Other autonomic effects observed after stimulating these areas were piloerection in monkeys (supracallosal portion of the anterior limbic cortex and olfactory tubercle) and slight pupillary dilatation (from the respiratory inhibitory areas); micturition and defecation occurred frequently on stimulating the pyriform cortex and amygdala.

Ablation of different parts of the rhinencephalic area produces certain autonomic effects. Thus ablation of area 13 is followed by vasodilatation and increase in skin temperature.[3] Bilateral ablation of the limbic gyrus provokes an increase in gastric motility.[4] Bilateral removal of area 13 causes hypermotility in the gastrointestinal tract and hyperphagia.

Somatic movements initiated in the neocortex are integrated with changes in the circulation and other visceral reactions. Thus stimulation

[1] Fulton, J. F., *Medicine*, **15**, 246, 1936.

[2] Papez, J. W., *Arch. Neurol. & Psychiat.*, **38**, 725, 1937.

[3] Kaada, B. R., *Acta physiol. Scandinav.*, **24**, Suppl. 83, 1951; Hess, W. R., *et al.*, *Helvet. physiol. pharmacol. acta.*, **9**, 101, 1951.

[4] Bailey, P., and W. H. Sweet, *J. Neurophysiol.*, **3**, 276, 1940; Smith, W. K., *J. Neurophysiol.*, **8**, 241, 1945; Kaada, B. R., K. H. Pribram, and J. A. Epstein, *J. Neurophysiol.*, **12**, 347, 1949; Sachs, E., S. J. Brendler, and J. F. Fulton, *Brain*, **72**, 227, 1949.

[1] Rioch, D. McK., and C. Bremmer, *J. Comp. Neurol.*, **68**, 491, 1938.

[2] Kaada, *loc. cit.*

[3] Delgado, J. M. R., and R. B. Livingston, *J. Neurophysiol.*, **11**, 39, 1948.

[4] Babkin, B. P., and W. C. Kite, *J. Neurophysiol.*, **13**, 335, 1950.

of area 6 in the frontal lobe provokes tachycardia, hypertension, and vasoconstriction in the splanchnic area and the skin, so that the circulating blood is shunted toward the muscles, brain, and lungs. Stimulation of neighboring areas may provoke vasodilatation and a fall in blood pressure. The pressor and depressor areas frequently overlap. Stimulation of the motor cortex of the dog between the sulcus cruciatus and a more caudal sulcus homologous to the Rolandic sulcus provokes vasodilatation in the limb muscles, through activation of cholinergic vasodilators, and vasoconstriction in the skin and splanchnic area; the blood flow through the muscles is thus increased. The same effects are obtained by stimulation of the supraoptic region near the mid-line in the cat and dog.[1] Stimulation of areas 6 and 8 in the frontal cortex is followed by augmentation of gastrointestinal movements,[2] though in some cases inhibitory effects are observed.[3] Stimulation of the cortical centers that provoke movements of the tongue also causes salivary secretion, and certain movements of the hand induced by cortical stimulation are accompanied by sweat secretion, which increases sensitiveness to touch. Stimulation of the frontal eye fields provokes not only movements of the eye muscles but also dilatation of the pupil, which is suppressed by removal of the superior cervical ganglion.[4] Another cortical center that provokes dilatation of the pupil has been found in the cat[5] in the medial aspect of the hemispheres, above the corpus callosum and anterior to the cruciate junction. Constriction of the pupil can be obtained by stimulation of the preoccipital area. Peripheral sympathetic and parasympathetic denervation modifies these responses; thus, removal of the stellate ganglia suppresses tachycardia following cortical stimulation, and section of the vagi increases it.

In all these cases somatic movements are integrated with adequate visceral reactions. Interconnections between somatic and visceral motor innervation are therefore established at all levels of integration.[1]

[1] ELIASSON, S., B. FOLKOV, P. LINDGREN, and B. UVNÄS, *Acta physiol. Scandinav.*, **23**, 333, 1951; **27**, 18, 1952.

[2] WATTS, J. W., and J. F. FULTON, *New England J. Med.*, **210**, 883, 1934; DAVEY, L. M., B. R. KAADA, and J. F. FULTON, *Research Publ., A. Nerv. & Ment. Dis.*, **29**, 619, 1950.

[3] SHEEHAN, D., *J. Physiol.*, **83**, 177, 1934.

[4] WARD, A. A., and H. L. REED, *J. Neurophysiol.*, **9**, 329, 1946.

[5] SIEBENS, A. A., and C. N. WOOLSEY, *Federation Proc.*, **5**, 95, 1946.

Vasomotor responses are disturbed by certain cortical lesions.[2] It is a well-known fact that in cases of hemiplegia there is at first an increase in the temperature of the skin on the paralyzed limbs, due to vasodilatation, and that later the paralyzed limbs are colder than the contralateral limbs, owing to vasoconstriction. Experimental destruction of the frontal cortex disturbs vasomotor reactions. Thus in the dog cold does not provoke vasoconstriction, and in the monkey heat does not provoke vasodilatation.

The reactions for regulating body temperature are sluggish in animals in which the frontal cortex has been destroyed, but once established they have a tendency to persist and outlast their need. The "psychogalvanic" reflex (decrease in the electrical resistance of the skin when certain auditory, visual, or tactile stimuli are applied, owing to an increase in sweat secretion) is integrated in the frontal cortex, since it is abolished by destruction of area 6. Stimulation of this area, and of the anterior part of the temporal lobe, provokes sweat secretion in the cat, even after the hypothalamus has been destroyed; the impulses travel along the frontopontine and temporopontine pathways.

Visceral functions are integrated at different levels, in a way similar to that of somatic activity. Simple local reactions are integrated by means of segmentary spinal reflexes. More complex patterns, such as the regulation of blood pressure and vomiting (which has also somatic components), are integrated in the medulla. Still more complex synergies, such as temperature regulation, are integrated at a hypothalamic level. Finally there is also a cortical level of integration at which visceral and somatic responses are coordinated.

PHYSIOLOGIC SIGNIFICANCE OF VISCERAL INNERVATION

Most organs function in apparently normal condition after complete denervation. Thus, gastrointestinal motility and secretion suffer only transitory disturbances after denervation. The salivary glands, however, are an exception

[1] KENNARD, M. A., *Psychosom. Med.*, **9**, 29, 1947.

[2] PINKSTON, J. O., P. BARD, and D. McK. RIOCH, *Am. J. Physiol.*, **109**, 515, 1934; PINKSTON, J. O., and D. McK. RIOCH, *Am. J. Physiol.*, **121**, 49, 1938.

to this rule, and salivary secretion is abolished permanently after the secretory nerves have been cut. The muscles of the blood vessels and orbital smooth-muscle structures soon recover their tonus after denervation. Likewise completely denervated or grafted endocrine glands, with the exception of the adrenal medulla and the pars nervosa of the hypophysis, function normally.

Moreover both vagi can be cut[1] and the whole sympathetic chain can be removed on both sides, from the superior cervical to the coccygeal ganglion, together with the solar plexus, with survival of the subjects.[2] Totally sympathectomized cats, dogs, and monkeys behave apparently normally in quiet conditions. No changes are observed in somatic motility, in growth, or in digestive, nutritive, and sexual functions (estrus, fertilization, pregnancy, parturition, and lactation). There are no apparent changes in psychic behavior; the temperament of the animal remains as before the operation, whether it was docile and affectionate or surly and aggressive. Somatic expression of emotion is not altered. A mother cat will defend her kittens from a dog by arching her back, spreading her claws, and hissing, just as a normal cat does; but her hairs do not stand on end, the heart rate does not increase, and there is no hyperglycemia; *i.e.*, visceral reactions are absent.

The inferiority of sympathectomized animals as compared to normal ones is evident when they are placed in conditions of emergency or stress. For example, if a sympathectomized cat is placed in a room at 0 to 5°C., its rectal temperature falls one or two degrees, owing to the absence of reflexes (piloerection, cutaneous vasoconstriction) that diminish the loss of heat. The BMR of these animals is within the normal range, although it is low, and heat production rises when they are placed in a cold environment, by an increase in muscle tonus and shivering, as in normal animals.[3] Another instance of the limited capacity to resist stress

[1] This should be done below the emergence of the inferior laryngeal nerve, in order to avoid laryngeal paralysis and choking. Complete vagotomy in man is followed by disturbances in gastric function.

[2] Cannon, W. B., J. T. Lewis, and J. W. Britton, *Boston Med. & Surg. J.*, **197**, 574, 1927.

[3] Cannon, W. B., H. F. Newton, E. M. Bright, V. Menkin, and R. M. Moore, *Am. J. Physiol.*, **89**, 84, 1929.

is the difficulty with which they recover a normal blood sugar after hypoglycemia, whence their increased sensitiveness to insulin. There are, however, differences between species. Thus, sympathectomized cats are hypersensitive to anoxia, and their capacity to perform physical exercise is reduced; sympathectomized dogs resist anoxia and cold better than cats and can perform physical exercise fairly well.[1]

Visceral innervation is not, therefore, indispensable for life, but it is not without physiologic significance. Careful study of a denervated organ shows that it does not respond normally to functional stress. For example, a grafted (therefore completely denervated) pancreas secretes insulin in greater or lesser amounts according to the level of the blood sugar (more when it is high, less when it is low); but if the rate of recovery of the normal blood-sugar level after hyperglycemia or hypoglycemia in normal animals is compared with that in animals with a grafted pancreas, it will be seen that the former recover the normal level faster than the latter.[2] Many other examples of delayed response or inaccurate adjustment to conditions of stress could be given.

In emergencies that involve the whole body, hyperactivity of the sympathetic is a prominent feature. The architecture of this system is such that general reactions are facilitated, although the mechanism for localized response also exists. In strenuous muscular exercise, asphyxia, intense emotion, and other conditions that require mobilization of all the resources of the organism ("fright, flight, or fight" pattern of behavior) there are tachycardia, vasoconstriction in the splanchnic area and the skin, evacuation of blood depots (spleen, skin), increase in circulating blood volume, and shunting of the blood from the abdominal viscera to the locomotor apparatus (muscles, lungs, brain). The bronchi are dilated, and the red cell count increases; liver glycogen is converted to glucose, and the blood sugar rises. The sympathetic plays a major part in this type of response, and its action is reinforced by an increase in adrenaline secretion. The body reacts to generalized emergencies by mobilizing its reserves and increasing the expenditure of energy.

[1] McDonough, F. K., *Am. J. Physiol.*, **125**, 530, 1939.

[2] Houssay, B. A., J. T. Lewis, and V. G. Foglia, *Rev. Soc. argent. de biol.*, **4**, 859, 1928; **5**, 1, 1929.

The parasympathetic becomes active in "localized emergencies," *i.e.*, stress applied to a single organ or group of organs. The architecture of this system is well adapted to this purpose. The ganglia are situated near the innervated viscera, and they serve a relatively limited area. The chemical mediator (acetylcholine) is rapidly inactivated after it has been released, so that it cannot exert its effects on distant organs, as is the case with the chemical mediator (sympathin) of the sympathetic. For example, the ingestion of food creates an "emergency" in the digestive tract, and the glands and muscle of this tract are brought into activity by parasympathetic impulses. The parasympathetic is usually active when the body, especially the locomotor apparatus, is at rest, and the result of its activity is the accumulation of energy.

Visceral innervation has not only an intermittent activity in moments of stress, but also a continuous or "tonic" action. Tonic impulses arise mainly, although not exclusively, in reflexes from the receptors in the aorta and the carotid sinus. Changes in blood pressure, in oxygen and CO_2 pressure, in acid-base equilibrium, and many other factors (see Chap. 17) act on these vascular receptors and cause a reflex increase or decrease of sympathetic and parasympathetic impulses, which tend to restore the normal equilibrium. For example, a rise in blood pressure stimulates the stretch receptors in the aortic wall and the carotid sinus, thus provoking reflex stimulation of the vagus and inhibition of the cardioaccelerator fibers in the sympathetic, so that bradycardia results; at the same time the vasoconstrictors are inhibited, and the secretion of adrenaline diminishes. All these reactions tend to lower the blood pressure and restore it to the normal level.

The visceral nervous system contributes to the rapid and accurate adaptation of the organism in changing circumstances. It is, therefore, an important mechanism of "homeostasis,"[1] helping to maintain that fixity of the *milieu intérieur* which, as Claude Bernard so clearly stated, is a condition of free life.

[1] CANNON, W. B., *Physiol. Rev.*, **9**, 399, 1929.

BIBLIOGRAPHY

BROWN, G. L., Transmission at Nerve Ending by Acetylcholine, *Physiol. Rev.*, **17**, 485, 1937.

BURN, J. H., Relation of Motor and Inhibitory Effects of Local Hormones, *Physiol. Rev.*, **30**, 177, 1950.

CANNON, W. B.: "Bodily Changes in Pain, Hunger, Fear and Rage," 2d ed., Appleton-Century-Crofts, New York, 1929.

COLLE, J., P. DELL, and H. GASTAUT, Corrélations entre le système nerveux végétatif et le système de la vie de relation, *J. Physiol.*, **44**, 415, 431, and 471, 1952.

DALE, H. H., Transmission of Nervous Effectors by Acetylcholine, *Harvey Lect.*, 1937.

GASKELL, W. H., "The Involuntary Nervous System," Longmans, New York, 1916.

HARE, K., The Visceral Functions of the Nervous System, *Ann. Rev. Physiol.*, **8**, 375, 1946.

"The Hypothalamus," *Proc. Assoc. Res. Nerv. Ment. Dis.*, **20**, 1940.

INGRAM, W. R., The Visceral Functions of the Nervous System, *Ann. Rev., Physiol.*, **9**, 163, 1947.

KAADA, B. R., Somato-motor, Autonomic, and Electrocorticographic Responses to Electrical Stimulation of Rhinencephalic and Other Structures in Primates, Cat and Dog, *Acta physiol. Scandinav.*, **24**, Suppl. 83, 1951.

KUNTZ, A., "The Autonomic Nervous System," 3d ed., Lea & Febiger, Philadelphia, 1945.

KUNTZ, A., Visceral Functions of the Nervous System, *Ann. Rev. Physiol.*, **14**, 409, 1952

LANGLEY, J. N., "The Autonomic Nervous System," Part 1, Cambridge, New York, 1921.

LE GROS CLARK, W. E., J. BEATTIE, G. RIDDOCH, and N. M. DOTT, "The Hypothalamus," Oliver & Boyd, Edinburgh and London, 1938.

LIVINGSTON, R. B., Visceral Functions of the Nervous System, *Ann. Rev. Physiol.*, **12**, 445, 1950.

MÜLLER, L. R., "Lebensnerven und Lebenstriebe," 3d ed., Berlin, 1931.

PI SUÑER, A., "Sistema Neurovegetativo," Unión Tipográfica Editorial Hispano-americana, Mexico, 1947.

ROSENBLUETH, A., "The Transmission of Nerve Impulses at Neuroeffector Junctions and Peripheral Synapses," Wiley, New York, 1950.

WANG, S. C., Visceral Functions of the Nervous System, *Ann. Rev. Physiol.*, **15**, 329, 1953.

WHITE, J. C., and R. H. SMITHWICK, "The Autonomic Nervous System. Anatomy, Physiology, and Surgical Application," Macmillan, New York, 1945.

WHITTAKER, V. P., Specificity, Mode of Action and Distribution of Cholesterinases, *Physiol. Rev.*, **31**, 312, 1951.

YOUMANS, W. B., Visceral Functions of the Nervous System, *Ann. Rev. Physiol.*, **11**, 139, 1949.

CHAPTER 85

The Secretion of the Adrenal Medulla

THE SECRETION OF the adrenal medulla, adrenaline (epinephrine) and noradrenaline (norepinephrine), produces the same effects as stimulation of the adrenergic sympathetic. The nervous and humoral mechanisms that activate sympathetic effectors have a common origin in the embryo and they form a functional unit, the sympathicoadrenal system. For this reason medulliadrenal secretion is discussed in this section rather than in Sec. VI, with the functions of the adrenal cortex and other internal secretions.

Embryonic origin of the adrenal glands. The cortical tissue arises from the mesenchyma, in the part of the celomic epithelium situated near the site of origin of the kidneys and the sexual glands. At an early stage of development many buds appear. Some are reabsorbed; others conglomerate and form the adrenal cortex; a few remain isolated and constitute accessory corticoadrenal glands. These are more frequently found in some species, such as the rat and rabbit, than in others, such as man, dog, cat, and guinea pig. They are usually situated near the gonads, in the epididymis or the broad ligament, and in the retroperitoneal tissue, along the course taken by the gonad in its fetal migration.

The tissue of the medulla arises in the ectoderm, in cells that form part of the buds for the spinal ganglia and later develop in one of two ways. Some of them, the sympathoblasts, are converted into sympathetic ganglion cells; others, the pheochromoblasts, take on a glandular aspect and form the adrenal medulla. The latter give typical histochemical reactions, because of the substance from which adrenaline is made: they stain blue with ferric chloride, and a dark brown with chromates. The latter reaction has given them the name of chromaffin cells. In the early stages of development many buds of chromaffin cells appear; some are reabsorbed, while others conglomerate into the adrenal medulla. A few groups remain isolated and form the accessory chromaffin organs found in the retroperitoneal tissue (Zuckerkandl's body) and near the sympathetic ganglia (paraganglia).

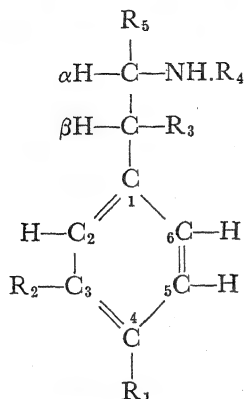
The adrenal is well provided with blood vessels, and has a very active circulation.

Chemical structure of adrenaline. Oliver and Schäfer[1] were the first to report that adrenal extracts produce a rise in blood pressure as a result of vasoconstriction. The extract exerts its effects on peripheral structures and is therefore active in pithed animals and on denervated organs. This fundamental observation led to the discovery of adrenaline. Abel[2] and others purified the extracts, and in 1901 Takamine[3] and Aldrich[4] obtained the pure substance and established its chemical structure. Later Stolz and then Dakin[5] prepared it by synthesis. Adrenaline is the characteristic component of chromaffin cells, which are identified by the histochemical reactions of adrenaline or the pharmacologic effects of their extract.

Adrenaline is α-methylamine-β-3,4;dioxyphenylethanol. The hydroxyl groups on C^3 and C^4 of the benzene ring place it among the cathecols; C^1 of the benzene ring is joined to

[1] OLIVER, G., and E. A. SCHÄFER, *J. Physiol.*, **16**, 1, 1894; **18**, 230, 1895.
[2] ABEL, J. J., *Am. J. Physiol.*, **2**, 13, 1899.
[3] TAKAMINE, J., *Am. J. Pharm.*, **73**, 523, 1901; *J. Physiol.*, **29**, 27, 1901.
[4] ALDRICH, I. B., *J. Physiol.*, **29**, 457, 1901.
[5] DAKIN, H. E., *J. Physiol.*, **32**, 32, 1905.

Cβ of a side chain, which has a secondary alcoholic OH; Cα has a methylated amine.

```
            R5
            |
      αH—C—NH.R4
            |
      βH—C—R3
            |
            C
          // 1 \
   H—C2        6C—H
      |          ||
   R2—C3       5C—H
          \\ 4 /
            C
            |
            R1
```

Key structural formula of the most common sympathomimetic substances (Hug)

In the benzene ring and the side chain, different chemical groups can be substituted for hydrogen. $C^3(R^2)$ and $C^4(R^1)$ in the ring, $C\beta(R^3)$ and $C\alpha(R^5)$ in the side chain, and the amine $N(R^4)$ are the most frequent sites of substitution. Thus in adrenaline (epinephrine) OH is substituted for three hydrogen atoms, one each at $C^4(R^1)$, $C^3(R^2)$, and $C\beta(R^3)$, and one hydrogen atom of the amine (R^4) by CH_3.

Table 117*

	R^1	R^2	R^3	R^4	R^5
Adrenaline, epinephrine	OH	OH	OH	CH_3	
Kephrine, adrenalone	OH	OH	O	CH_3	
Epinine	OH	OH	....	CH_3	
Noradrenaline, arterenol	OH	OH	OH		
Isopropylnoradrenaline	OH	OH	OH	$CH(CH_3)_2$	
Ephedrine	....	....	OH	CH_3	CH_3
Norephedrine	....	....	OH	....	CH_3
Cobefrine, corbasil	OH	OH	OH	....	CH_3
Benzedrine, amphetamine	....	....	....	CH_3	
Metamphetamine	....	....	....	CH_3	CH_3
Tyramine	OH				

* For a list of synonyms see *J. A. M. A.*, **109,** 2065, 1937, and for other sympathomimetic drugs see TAINTER, M. L., *Arch. internat. de pharmacodyn. et de thérap.*, **46,** 191, 1933.

Adrenaline forms part of the group of substances known as alkaloids, and as such has some properties in common with these; *e.g.*, it is only slightly soluble in water but is readily soluble in organic solvents, and the saturated solution is alkaline; with acids it forms soluble salts. As it is a cathecol, it acts as a reducing agent and gives typical color reactions. It has an asymmetric carbon, that of the secondary alcohol in the side chain, which endows it with optical activity. The gland secretes the levorotatory form, which has a specific rotation of $-53°$ when in a 1 per cent solution in 0.1 N H_2SO_4 at 20°C. The racemic form has been prepared by synthesis and the dextrorotatory form has then been separated from it. Natural adrenaline is less stable but has from 10 to 20 times the pharmacodynamic activity of D-adrenaline. The racemic form is approximately one-half as active as L-adrenaline. The relative activity of the optical isomers varies in different effectors.

The normal adrenal medulla of most animals, including man, contains not only L-adrenaline but also L-noradrenaline, *i.e.*, the nonmethylated form.[1] The relative amounts of the amines vary. There is very little or no noradrenaline in rabbits and guinea pigs; in cattle, cats, and dogs, adrenaline constitutes about 80 per cent of the total; in the toad there is 2.5 to 5 mg./gm., made up of approximately equal parts of both amines.[2] In whales there is 4 to 5 mg./gm. of medulliadrenal tissue, 70 to 100 per cent of which is noradrenaline.[3] Chromaffin-cell tumors removed from human subjects have been found to contain large amounts of noradrenaline (5.3 to 6.3 mg./gm. of tumor tissue) and smaller quantities of adrenaline (0.32 to 3.7 mg./gm).[4] In other cases adrenaline was the predominant amine.[5]

Denervation of the adrenals slightly increases the catechol content but does not change the

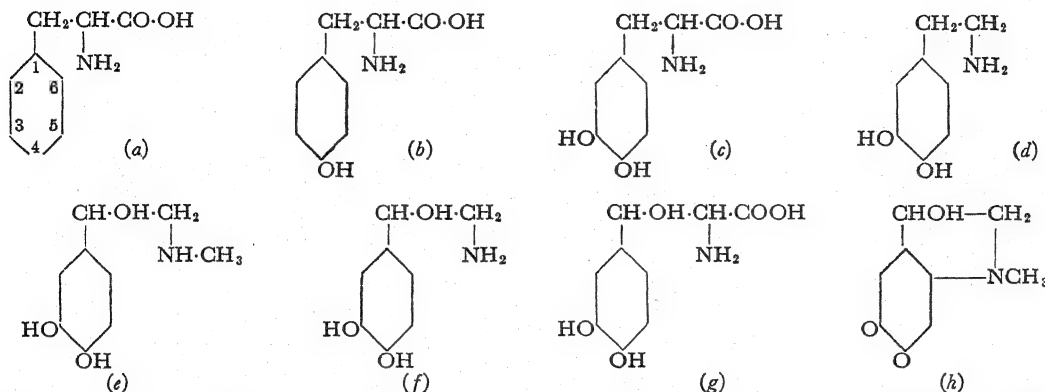
[1] HOLTZ, P., *Verhandl. deutsch. physiol. Ges., Frankfurt,* 1948; BÜLBRING, E., and J. H. BURN, *Nature, London,* **163,** 363, 1949; EULER, U. S. VON, and U. HAMBERG, *Nature, London,* **163,** 642, 1949; GOLDENBERG, M., *et al., Science,* **109,** 534, 1949; HOLTZ, P., and SCHÜMANN, *Nature, London,* **165,** 683, 1950; BERGSTRÖM, S., U. S. VON EULER, and U. HAMBERG, *Acta physiol. Scandinav.,* **20,** 101, 1950; HÖKFELT, B., and J. MCLEAN, *Acta physiol. Scandinav.,* **21,** 258, 1951.

[2] HOUSSAY, B. A., R. GERSCHMAN, and C. E. RAPELA, *Rev. Soc. argent. de biol.,* **26,** 29, 1950.

[3] BURN, J. H., *et al., J. Physiol.,* **113,** 123, 1951.

[4] HOLTON, P., *Nature, London,* **163,** 217, 1949.

[5] PITCAIRN, D. M., and W. B. YOUMANS, *Circulation,* **2,** 505, 1950.

relative amounts of adrenaline and noradrenaline. Prolonged stimulation of the adrenal, *e.g.*, by insulin or urethane, decreases the adrenaline content but increases the relative, and in some cases the absolute, amount of noradrenaline.[1] oxidized into a cathecol before it is decarboxylated. Decarboxylation takes place before the side chain is modified; after introduction of an *N*-methyl group, or a hydroxyl group in the position β, the compounds are no longer

Possible steps in the formation of adrenaline: *a*, phenylalanine; *b*, tyrosine; *c*, dihydroxyphenylalanine; *d*, dihydroxyphenylethylamine; *e*, adrenaline; *f*, noradrenaline; *g*, dihydroxyphenylserine; *h*, adrenochrome.

The formation of adrenaline. Phenylalanine (*a*), labeled with C^{14} or tritium, has been shown to be converted *in vivo* into tyrosine (*b*), 3, 4-dihydroxyphenylalanine (*c*), dihydroxyphenylethylamine (*d*) and eventually into adrenaline (*e*).[2] The first stage is the oxidation of the ring; phenylalanine is oxidized to tyrosine. The conversion of L-phenylalanine into L-tyrosine is catalyzed by a highly specific enzyme system found in rat liver, which is active in the presence of DPN (diphosphopyridine nucleotide) and oxygen.[3] Further oxidation converts tyrosine into 3,4-dihydroxyphenylalanine (DOPA). Oxidation of the ring precedes oxidation of the side chain; hydroxytyramine is converted into adrenaline by adrenal tissue, but phenylhydroxyethylamine is not.[4] The second stage is decarboxylation, catalyzed by DOPA decarboxylase. This enzyme acts on dihydroxyphenylalanine[5] but not on tyrosine,[6] *i.e.*, the ring is catalyzed by DOPA decarboxylase. The last stage in the process is *N*-methylation, *i.e.*, the conversion of noradrenaline (*f*) into adrenaline by a process of transmethylation in which methionine acts as methyldonor.[1] It is also possible that noradrenaline is formed by decarboxylation of dihydroxyphenylserine (*g*) which can be considered as the carboxylic acid of noradrenaline; this process has been demonstrated *in vitro*, catalyzed by guinea-pig liver and kidney extracts in anaerobic conditions.[2] Methylation of noradrenaline by the adrenals has been demonstrated *in vitro* by incubating it with minced adrenal tissue, and *in vivo* by perfusing the adrenal glands of dogs with heparinized blood containing noradrenaline; in both cases adrenaline increased and noradrenaline decreased.[3] Conversion of noradrenaline into adrenaline is also performed by other tissues.[4]

Inactivation of adrenaline. Adrenaline in solution undergoes a slow process of auto-

[1] Outschoorn, A. S., *Nature, London*, **167**, 722, 1951; Hökfelt, and McLean, *loc. cit.*

[2] Gurin, S., and A. M. Delluva, *J. Biol. Chem.*, **170**, 545, 1947.

[3] Udenfriend, S., and J. R. Cooper, *J. Biol. Chem.*, **194**, 503, 1952.

[4] Holtz, P., and Kroneberg, *Arch. f. exper. Path. u. Pharmakol.*, **206**, 150, 1949.

[5] Holtz, P., R. Heise, and K. Lüdke, *Arch. f. exper. Path. u. Pharmakol.*, **191**, 87, 1938.

[6] Meta- and ortho-tyrosine, substances which are not normal constituents of mammalian tissues, are also substrates for DOPA decarboxylase (Blaschko *et al.*, *J. Physiol.*, **108**, 427, 1949; **110**, 482, 1949).

[1] Keller, E. B., R. A. Boissonas, and V. du Vigneaud, *J. Biol. Chem.*, **183**, 627, 1950.

[2] Blaschko, H., J. H. Burn, and H. Langermann, *Brit. J. Pharmacol.*, **5**, 431, 1950.

[3] Bülbring, E., *Brit. J. Pharmacol.*, **4**, 234, 1949; Bülbring, E., and J. H. Burn, *Brit. J. Pharmacol.*, **4**, 245, 1949.

[4] Bacq, Z. M., *Science*, **108**, 135, 1948.

oxidation; adrenochrome (*h*), which gives a pink color to the solution, is formed, and part of the pharmacodynamic activity is lost. This process is accelerated *in vitro* by the cytochrome–cytochrome oxidase system and by other enzymes found in plants, such as catecholoxidase and tyrosinase. Prolonged oxidation gives rise to melanin. In the first stages of oxidation it is probable that a very unstable orthoquinone of adrenaline is formed.

Adrenaline is inactivated in the organism by an enzymatic mechanism probably similar to those acting *in vitro*, but no satisfactory proofs have been given that a particular system is the principal or only one responsible.

When large doses of adrenaline are taken by mouth, sulfoconjugated compounds of adrenaline increase in the urine.[1] The intestine and the liver are important sites of sulfoconjugation, and it is possible that on passing through these organs adrenaline is inactivated by this process, but sulfoconjugation is of doubtful importance as a means of adrenaline inactivation in muscle.

In blood and in certain tissues adrenaline is more stable than in watery solutions. This stability is due to the presence of protecting substances acting as reducing agents, among which there are several amino acids, glutathione, and ascorbic acid. The adrenal gland contains large quantities of glutathione and ascorbic acid, but it is not known whether these substances play a part in the stabilization of adrenaline secreted by the gland.

The action of adrenaline on the organism. Langley,[2] studying the effects of adrenal extracts, and Elliott,[3] studying those of adrenaline, demonstrated that the active principle of the adrenal medulla has the same effect on the organism as stimulation of the sympathetic nerves. There are a few exceptions to this rule. In some cases the sympathetic fibers are cholinergic, *e.g.*, the innervation of the sweat glands in some species, and some of the sympathetic vasodilators. There are also adrenergic fibers in the parasympathetic (see Chap. 84).

Adrenaline acts on what Langley called the "receptive substance" of the effector. The chemical nature of this substance and the way

[1] RICHTER, D., *J. Physiol.*, **98**, 361, 1940.
[2] LANGLEY, J. W., *J. Physiol.*, **27**, 237, 1901.
[3] ELLIOTT, T. R., *J. Physiol.*, **32**, 401, 1905.

it reacts with adrenaline are still unknown. Denervation not only does not suppress the effects of adrenaline but sensitizes the tissue to these effects.[1] Hypersensitiveness to adrenaline and to sympathetic stimulation is produced in some cases, *e.g.*, the salivary glands, by suppressing parasympathetic impulses by denervation or treatment with atropine for several days.[2] Hypersensitiveness is established gradually after denervation and reaches its maximum in a few days. It can be produced by section of the preganglionic fibers, but is more marked after postganglionic denervation.

Adrenaline must be in a certain minimum concentration to produce its effects; below this minimum either there is no response or even a contrary effect is obtained. Thus, adrenaline provokes hypertension due to vasoconstriction, but very small doses provoke vasodilatation and hypotension. On increasing the dose of adrenaline, the effect increases up to a maximum. For example, in anesthetized dogs, with both vagi cut so as to suppress the cardiomoderator reflex, intravenous injection of 0.2 μg per kg. provokes a small rise in blood pressure; on increasing the dose to between 0.5 and 0.7 μg per kg., there is a considerable increase in the hypertension, but a further increase in the dose provokes very little increase in the effect. The relation between dose and response has been determined in several effectors (contraction of arterial strips, contraction of the nictitating membrane, contraction or relaxation of the uterus, rate of the denervated heart, blood pressure, etc.). The results correspond to rectangular hyperbolas (Fig. 493) when the response of the effector is not modified by the interference of homeostatic mechanisms.[3]

Sensitiveness to adrenaline varies according to the effector concerned. In the dog the denervated heart responds with tachycardia to very small doses, but the blood pressure is an even more sensitive indicator of adrenaline. Contraction of the spleen is obtained with smaller doses than those necessary for constriction of the kidney, and the vasomotors of

[1] CANNON, W. B., and A. ROSENBLUETH, "Supersensitivity of Denervated Structures," Macmillan, New York, 1949.
[2] EMMELIN, N., and A. MUREN, *Nature, London*, **166**, 610, 1950.
[3] ROSENBLUETH, A., *Am. J. Physiol.*, **101**, 149, 1932.

the limbs are less sensitive. Relatively large doses are necessary in order to produce glycogenolysis and hyperglycemia.

The different animal species are not equally sensitive to adrenaline. In man continuous intravenous injection of 0.05 μg per kg. per min. increases the blood pressure; in the dog, 10 times this dose is needed to obtain a similar response, and in the rat, 20 times. Subcutaneous injection of adrenaline produces local vasoconstriction and thus retards the absorption of the drug. Perhaps it is for this reason that in the dog and the rat it has no effect on the circulation when injected subcutaneously, while man, being more sensitive, responds with hypertension, tachycardia, increased cardiac output, and a larger respiratory minute volume.

An increase in sensitiveness to adrenaline can be obtained in several ways: (*a*) by denervation, as has already been mentioned; (*b*) by suppressing the regulatory mechanisms that oppose the effects of adrenaline; *e.g.*, if the depressor fibers are cut, reflex bradycardia, provoked by dilatation of the aorta and the carotid sinus caused by adrenaline hypertension, is suppressed; (*c*) by the administration of certain drugs, such as cocaine. Thus, dogs can be sensitized to adrenaline by suppressing the cardiomoderator mechanism (either by cutting the vagi or injecting atropine) and injecting cocaine, following which they will respond with an increase in blood pressure to the intravenous injection of 0.005 μg per kg., and in some cases to even smaller doses. Hypothermia, on the other hand, decreases sensitiveness to adrenaline.[1]

The effects of adrenaline are of very short duration. The increase in blood pressure reaches its maximum in 1 or 2 min., returns to the initial level in 5 or 6 min., and then falls below this level. The secondary hypotension has been attributed to the formation of adrenoxine,[2] a product of oxidation of adrenaline, which has an inhibitory effect on vasoconstrictors. Another explanation given is that adrenaline concentration drops to an inhibitory level. The briefness of the action of adrenaline may be due in part to its destruction by the tissues, but it has been shown that effectors cease to respond after a time, even when there are still large amounts of adrenaline in the blood. Thus, if blood is drawn from a dog more than 30 min. after the effects of injecting a large dose (10 mg.) have ceased and is injected into another dog, it provokes hypertension and other effects of adrenaline.[1] Adrenaline can, therefore, be considered a "potential drug,"[2] *i.e.*, a substance that produces an effect depending not so much on the total concentration in the tissues as on the ratio between the concentrations within and without the cells on which it acts. When the internal and external concentrations are equalized, there are no longer any effects, although large amounts of the drug may still be found in the blood.

FIG. 493. Concentration-action curve of adrenaline. Cat with pithed brain and denervated heart. Abscissa, dose of adrenaline (unit = 2.5 μg). Ordinate, maximal increase in frequency of heart rate in 15 sec. over the basal rate of 35 beats in 15 sec. (*Rosenblueth, A., Am. J. Physiol., vol. 101, p. 149, 1932.*)

Adrenaline and noradrenaline have remarkable effects on the circulation. They cause a rise in blood pressure, mainly due to vasoconstriction; the pressor effect of L-noradrenaline is approximately 60 per cent greater than that of L-adrenaline.[3] The effects on the blood vessels vary in different vascular territories: Adrenaline constricts the vessels of the splanchnic area and the skin. In small doses it dilates the blood vessels of the muscles; in large doses it constricts them. It has little or no effect on the pulmonary and

[1] VALLERY-RADOT, P., G. MAURIC, A. HOLTZER, A DOMART, and J. LEHMANT, *Compt. rend. Soc. de biol.*, **137**, 78, 1943.

[2] BACQ, Z. M., and P. HEIRMAN, *Ann. physiol. physicochim. biol.*, **14**, 476, 1938.

[1] COLLADOS-STORNI, M. P., thesis for M.D., University of Córdoba, 1943.

[2] TIFFENAU, M., Sur les poisons potentiels, Kongressberichte d. XVI Internat. Physiol. Kongr., Zurich, **1938.**

[3] TAINTER, M. L., B. F. TULLAR, and F. P. LUDUEÑA, *Science*, **107**, 39, 1948.

cerebral circulations, and it dilates the coronary arterioles. Noradrenaline has similar effects, but it has only a constrictor action on muscular blood vessels; in man peripheral resistance is increased by noradrenaline and lowered by adrenaline. Both amines provoke constriction of the spleen, thus increasing the circulating blood volume and the erythrocyte concentration. The heart rate and the strength of the heartbeat are increased, but in the intact animal the rise in blood pressure stimulates the receptors in the aorta and the carotid sinus, and reflex vagal bradycardia follows after the initial tachycardia. Adrenaline, therefore, accelerates the circulation and shunts the blood to the locomotor apparatus: muscles, central nervous system, heart and lungs, *i.e.*, establishes circulatory conditions favorable for the performance of physical effort.

Adrenaline provokes constriction of the capillaries, an effect which is often used in therapeutics with the object of reducing the circulation locally, or to counteract capillary dilatation, *e.g.*, in certain allergic conditions.

Adrenaline increases oxygen consumption by 15 to 30 per cent. The calorigenic effect of noradrenaline is about 10 per cent that of adrenaline and can be accounted for by the increase in lactic acid due to hypoxia caused by constriction of muscular blood vessels.[1]

Adrenaline produces hyperglycemia; noradrenaline has a small hyperglycemic effect; only 5 to 10 per cent that of adrenaline. Adrenaline hyperglycemia is due to:

1. An increase in glycogenolysis. The hepatic glucose output, determined in normal men by catheterization of the hepatic vein, increases considerably (noradrenaline produces only a slight increase).[2] Moreover adrenaline hyperglycemia is reduced if the hepatic glycogen stores have been depleted.
2. Inhibition of glucose consumption by the tissues, as shown by the arteriovenous difference in blood-glucose concentration after injecting adrenaline.[3] Adrenaline decreases glucose tolerance in eviscerated rats,[4] and inhibits glucose utilization by the diaphragm.

[1] LUNDHOLM, L., *Acta physiol. Scandinav.*, **21**, 195, 1951.
[2] BEARN, A. G., B. BILLING, and S. SHERLOCK, *J. Physiol.*, **115**, 430, 1951.
[3] SOMOGYI, M., *J. Biol. Chem.*, **186**, 513, 1950.
[4] INGLE, D. J., and J. E. NEZAMIS, *Endocrinology*, **46**, 14, 1950.

The glucose uptake by muscle extracts is not reduced by previous administration of adrenaline until adenosinetriphosphate concentration becomes the limiting factor. It seems that adrenaline acts on ATP, not on hexokinase.[1] Apparently when adrenaline is oxidized one of its products can bind phosphate,[2] thus interfering with the utilization of glucose (see Chap. 41).

Adrenaline accelerates recovery after fatigue and increases the twitch tension of nonfatigued striated muscle. This effect is associated with an increase in the resting potential and a decrease in the rate of potassium loss from muscle.[3]

The significance of adrenaline and noradrenaline in the transmission of sympathetic impulses has already been discussed (see Chap. 84). Adrenaline has many other effects such as mydriasis, bronchial dilatation, and inhibition of intestinal movements.

The action of noradrenaline on many effectors is equal to that of adrenaline. Some of its excitor effects are slightly greater than those of adrenaline, *e.g.*, the pressor effect, but others are less, *e.g.*, the action on the heart rate. Noradrenaline has much less inhibitor action on the arterioles and on the uterus (rat, virgin cat, pregnant guinea pig), but its inhibitory action on the intestine is greater than that of adrenaline (see "The effects of drugs on visceral innervation," Chap. 84).

Biological assay of adrenaline. Isolated effectors are often used in the assay of adrenaline. The following are some of the methods used most frequently:

1. A strip of rabbit intestine, preferably the duodenum, is submerged in Ringer's solution and the spontaneous rhythmic contractions are registered. On adding adrenaline the movements cease and the intestine relaxes.
2. The rate of contraction of the isolated heart of a frog or toad increases on adding adrenaline to the perfusion fluid.
3. The Läewen-Trendelenburg method consists in the perfusion of the hindquarters of a frog or toad

[1] COHEN, J. A., and D. M. NEEDHAM, *Biochim. et biophys. acta*, **6**, 141, 1950.
[2] CHAIX, P., J. CHAUVET, and J. JEZEQUIEL, *Biochim. et biophys. acta*, **4**, 471, 1950; **5**, 472, 1950.
[3] BROWN, G. L., M. GOFFART, and M. VIANNA DIAS, *J. Physiol.*, **111**, 184, 1950; GOFFART, M., and W. L. M. PERRY, *J. Physiol.*, **112**, 95, 1951.

by the introduction of a cannula into the abdominal aorta, through which the perfusion fluid is sent at a constant pressure. Another cannula is placed in the abdominal vein in order to collect and measure the fluid passing through the blood vessels. The rate of flow depends on the diameter of the vessels. Adrenaline provokes vasoconstriction and reduces the amount of fluid passing through the system.
4. The enucleated eye of a frog or toad, when submerged in a solution containing adrenaline, shows mydriasis, and the pupil takes on a rounded shape.
5. Contraction of the nictitating membrane, the rate of the denervated heart, the rise in blood pressure, vasoconstriction in a denervated limb or a rabbit's ear, and many other organic responses to adrenaline have been used in order to assay the adrenaline content of blood or tissue extracts. If the active substance is adrenaline, the effects will be potentiated by cocaine and suppressed by sympatholytic drugs, such as ergotamine, iohimbin, dibenamine, etc.

Sympathomimetic drugs. Several drugs, structurally related to adrenaline, have a similar pharmacodynamic effect. Their activity differs from that of adrenaline in one or more of the following ways: (*a*) they are less potent; (*b*) they have a greater effect on some effectors and a lesser effect on others; (*c*) they do not provoke some of the effects of adrenaline, *e.g.*, they have no inhibitory action; (*d*) they differ in the response to potentiating and sympatholytic drugs. Two of these drugs are of particular interest in medicine:

1. Ephedrine, obtained by Chen[1] from ma huang (*Ephedra vulgaris*), a Chinese medicinal plant. Its potency on the blood pressure is a hundred times less than that of adrenaline, but its effects are more lasting. It has greater stability and is active when given by mouth.
2. Benzedrine (amphetamine), which has a weak peripheral activity[2] and a powerful action on the central nervous system. It provokes a condition of wakefulness, with great lucidity, euphoria, and loquacity.[3] It can have harmful effects.

[1] CHEN, K. K., and C. F. SCHMIDT, *J. A. M. A.*, **87**, 936, 1926.
[2] Benzedrine has little activity on the periphery when given in systemic doses, but produces effects when applied locally.
[3] COUNCIL ON PHARMACY, Report on Benzedrine, *J. A. M. A.*, **109**, 2064, 1937.

THE SECRETION OF ADRENALINE

Almost a century ago, Vulpian[1] observed that a blue color is obtained when adrenal blood is mixed with ferric chloride; this effect is due to adrenaline secreted by the gland. Later, adrenal blood was shown to provoke hypertension, vasoconstriction, mydriasis, intestinal relaxation, etc.

Methods for measuring adrenal secretion. There are several methods for measuring the adrenaline secreted in different physiologic states. The following are among those most frequently used:

1. Adrenal blood is collected during a given time, either directly or in a pocket made out of the vena cava by tying off all its tributaries except the adrenal vein. The action of this blood is tested on an appropriate effector, *e.g.*, an isolated strip of duodenum, and its effect is compared with that of a known amount of adrenaline.[2]
2. The adrenal vein is clamped, and after a time, during which adrenal blood has accumulated, the clamp is removed. The effects on a denervated sympathetic effector, such as the nictitating membrane, the iris (mydriasis), the spleen (contraction), or the heart (acceleration), are observed. The same procedure is followed after denervating the adrenal or extirpating the adrenal medulla.
3. The adrenal vein is joined to the jugular vein of another animal, and the adrenal blood is thus diverted into the receptor (Fig. 103, page 194), on which the effects of adrenal discharge are observed by registering the variations of an appropriate sympathetic effector (blood pressure, denervated heart or spleen, blood sugar, etc.).[3]
4. The adrenal gland is perfused with heparinized blood, and the adrenal hormones in the blood leaving the adrenal are measured by chemical or biological methods.

Operative proceedings and certain anesthetics, *e.g.*, ether, increase adrenaline secretion. In all the methods, the adrenaline secreted is

[1] VULPIAN, A., *Compt. rend. Acad. de sc.*, **43**, 663, 1856.
[2] STEWART, J. N., and J. M. ROGOFF, *J. Pharmacol. & Exper. Therap.*, **8**, 479, 1916.
[3] TOURNADE, A., and M. CHABROL, *Compt. rend. Soc. de biol.*, **85**, 651, 1921.

not always collected without loss; therefore the figures given must be considered as only approximations to the true figures.

Adrenaline in the blood. There is a very small amount of adrenaline in arterial blood, and the methods available for assaying it are not reliable. Chemical methods give results far above the true amounts, because substances that are not adrenaline are measured at the same time. Biological methods have shown that the concentration of adrenaline in the blood of the general circulation is around 0.0001 to 0.01 μg per cent. Adrenaline in the blood of the adrenal vein dialyzes through collodion membranes; therefore it cannot be bound to a protein.[1] A lactilic ester of adrenaline has been extracted from adrenal tissue, but there is no proof that circulating adrenaline is esterified. According to Bülbring and Burn,[2] 20 to 80 per cent of the activity of adrenal blood collected during stimulation of the splanchnic nerve is due to noradrenaline.

The rate of secretion and destruction of adrenaline (or equivalent sympathomimetic substances) has been calculated from the results of continuous infusion of adrenaline into dogs in which the sympathetic system has been blocked by complete spinal anesthesia with procaine. The blood pressure in these animals falls from the normal 125 to 130 mm. Hg to 6 mm. Hg, but it can be raised and kept at a normal level by continuous infusion of adrenaline at a rate of 0.45 μg/kg./min. The total amount of adrenaline (or its equivalents) active in the body has been calculated to be 0.71 μg/kg. It is being eliminated or destroyed continuously at a rate directly proportional to the total amount.[3]

The urine of normal men contains adrenaline, noradrenaline, and hydroxytyramine. The daily output of adrenaline has been found to be 11.5 ± 6 μg and that of noradrenaline 29 ± 12.3 μg.[4]

Secretory innervation (Fig. 494). Stimulation of the splanchnic nerves provokes the discharge of as much as 100 μg of adrenaline. The amount of adrenaline released is conditioned by

[1] Lewis, J. T., and T. J. Combes, *Rev. Soc. argent. de biol.*, **14**, 565, 1938.

[2] Bülbring, E., and J. H. Burn, *Nature, London*, **163**, 363, 1949.

[3] Guyton, A. C., and W. M. Gillespie, *Am. J. Physiol.*, **165**, 319, 1951.

[4] Euler, U. S. von, and S. Hellner, *Acta physiol. Scandinav.*, **22**, 161, 1951.

1. Adrenaline stored in the gland. Thus glands containing an average of 1.09 mg./gm. secreted 4 μg/min. on splanchnic stimulation, and glands with an average of 2.3 mg./gm. secreted 9.3 μg per min.[1]
2. The concentration of adrenaline in the blood. High concentrations obtained by intra-arterial injection of adrenaline or adding adrenaline to the blood perfusing isolated adrenals, inhibit adrenaline secretion and the gland does not respond to splanchnic stimulation for several minutes. Later the response is greater than before adrenaline injection (facilitation).[2] Bülbring and her associates have found the response to splanchnic stimulation increases when adrenaline is added to the perfusion fluid up to an optimal concentration; at higher concentrations it diminishes progressively and may be completely inhibited.

The secretory fibers emerge from the spinal cord in the thoracic and the first lumbar roots and reach the gland by the major and minor splanchnic nerves. The nerve fibers ramify and innervate several cells, forming neuroglanduiar units, similar to motor units. Chromaffin cells are modified nerve cells, similar to ganglionic neurons, and they secrete adrenaline, *i.e.*, they are adrenergic, as are most of the postganglionic fibers of the sympathetic. The adrenal medulla is innervated by preganglionic cholinergic fibers. Excitation of the splanchnic nerves releases acetylcholine, and this substance provokes adrenaline secretion. The gland responds to drugs in the same way as a sympathetic ganglion does. For example, nicotine stimulates the secretion of the chromaffin cells but blocks the secretory impulses from the splanchnic nerve; atropine in small doses has no effect, but in large doses it temporarily blocks the splanchnic nerve.

Nerve centers of adrenal secretion. There is a *spinal center* (or centers) for adrenal secretion situated in the upper thoracic segments, which exerts a continuous, or tonic, action on the adrenal medulla. In anesthetized or decerebrate animals, the adrenal secretes from 0.1 to 1 μg adrenaline per kilogram of body weight per

[1] Bülbring, E., J. H. Burn, and F. J. de Elio, *J. Physiol.*, **107**, 223, 1948.

[2] King, E. E., and A. S. Marrazzi, *Am. J. Physiol.*, **171**, 612, 1952.

minute.[1] Transection of the cord at the level of the lower cervical segments does not suppress this "spontaneous" secretion, but if the section is made below the third thoracic segment, adrenaline secretion is reduced to a minimum. The center is bilateral; semisection of the spinal cord suppresses secretion by the ipsilateral gland, but does not disturb the secretion of the contralateral one.[2] Denervation of the gland does not suppress adrenaline secretion completely. By means of the adrenojugular transfusion method, with a sensitized receptor, a secretion of approximately 0.005 μg per kg. of body weight per minute has been demonstrated.[3] The denervated adrenal does not atrophy; the histologic aspect remains normal, and several months after denervation adrenaline secretion can be provoked by means of acetylcholine and other drugs that act directly on the chromaffin cells.[1]

There is a *secretory center in the pons and the medulla*. Electrical stimulation of the upper half of the floor of the fourth ventricle provokes adrenaline discharge. A section of the brain stem immediately caudal to the posterior colliculi does not suppress reflex adrenaline secretion produced by centripetal stimulation of an afferent nerve, but if the section is made 2 mm. caudal to the posterior colliculi, secretion does not occur; this reflex is therefore integrated in a center located between the two sections.[2] Stimulation of the *hypothalamus* also produces adrenaline discharge, but in smaller quantities than stimulation of the medulla.[3] Both these centers send their impulses to the gland through the splanchnic nerves, and section of these nerves suppresses the effects of stimulating the

FIG. 494. Adrenojugular transfusion. Upper tracing: arterial blood pressure of donor, registered by elastic manometer. Middle tracing: arterial blood pressure of receptor, registered by mercury manometer. Lower tracing: volume of denervated hind limb of receptor. 1, stimulation of left splanchnic nerve of donor; 2, stimulation of the medulla; 3, stimulation of the medulla after cutting the left splanchnic nerve. (*Houssay, B. A., and E. A. Molinelli, Rev. Asoc. méd. argent. (Soc. biol.), vol. 37, p. 235, 1924.*)

[1] STEWART and ROGOFF, *loc. cit.*

[2] STEWART, G. N., and J. M. ROGOFF, *Proc. Soc. Exper. Biol. & Med.*, **14**, 143, 1917; *Am. J. Physiol.*, **48**, 397, 1919; **51**, 484, 1920.

[3] LEWIS, J. T., and R. O. PRIETO, *Rev. Soc. argent. de biol.*, **14**, 555, 1938.

[1] SGROSSO, J., *Rev. Soc. argent. de biol.*, **11**, 139, 1935.

[2] CANNON, W. B., and D. RAPPORT, *Am. J. Physiol.*, **58**, 338, 1921.

[3] HOUSSAY, B. A., and E. A. MOLINELLI, *Rev. Soc. argent. de biol.*, **6**, 600, 1925.

centers (Fig. 494). Stimulation of the cerebral cortex provokes adrenal discharge only when it produces muscular activity; a localized center of adrenal secretion has not been found in the cerebral cortex. According to Rogoff and his associates,[1] there is *an inhibitory center for adrenaline secretion*, located in the region of the brain stem bounded by the optic chiasma and the superior colliculi.

Physiologic conditions in which adrenaline secretion increases. *Centripetal stimulation of an afferent nerve* (*e.g.*, the sciatic, the brachial, the vagus) or any stimulus sufficiently strong to provoke pain produces adrenaline discharge (reflex secretion).[2]

Muscular contraction produces metabolites, which pass into the blood and stimulate the adrenal nerve centers. Even small active movements, but not passive ones, increase adrenaline secretion.[3] Exercise may cause an increase in the adrenaline content of the blood of trained animals and men, but a decrease occurs in untrained subjects.[4]

Emotional states, whatever their nature (rage, fear, etc.), that are accompanied by sympathetic activity also produce adrenaline secretion in relation to their intensity.[5]

Cold applied on the skin or the digestive tract (by means of iced drinks) increases adrenaline secretion. Thus the calorigenic effect of cold is reinforced by that of adrenaline, which also contributes to diminish the loss of heat by constricting the cutaneous blood vessels. Animals with denervated adrenals when put into a cold environment shiver intensely, more markedly than normal animals; thus the lack of the calorigenic effect of adrenaline is compensated by vigorous muscular activity.

[1] ROGOFF, J. M., P. WASSERMAN, and E. NOLA NIXON, *Proc. Soc. Exper. Biol. & Med.*, **61**, 251, 1946.

[2] CANNON, W. B., and R. CARRASCO-FORMIGUERA, *Am. J. Physiol.*, **61**, 215, 1922; HOUSSAY, B. A., and E. A. MOLINELLI, *Rev. Asoc. méd. argent.* (Soc. Biol.), 37, 327, 1934; *Rev. Soc. argent. de biol.*, **6**, 102, 1925.

[3] CANNON, W. B., J. R. LINTON, and R. R. LINTON, *Am. J. Physiol.*, **71**, 153, 1924; HOUSSAY, B. A., and E. A. MOLINELLI, *Rev. Soc. argent. de biol.*, **6**, 125, 1925; CANNON, W. B., and S. W. BRITTON, *Am. J. Physiol.*, **79**, 433, 1927.

[4] LEHMANN, G., and H. F. MICHAELIS, *Arbeitsphysiol.*, **12**, 218, 1943; HARMAN, N., H. F. MICHAELIS, and A. SZÁKALL, *Arbeitsphysiol.*, **13**, 57, 1944.

[5] CANNON, W. B "Bodily Changes in Pain, Hunger, Fear and Rage," 2d ed., Appleton-Century-Crofts, New York, 1929.

Hypotension produced by hemorrhage, by stimulation of the vagus, or in any other way provokes adrenaline discharge. Adrenaline secretion is regulated by the arterial blood pressure through the receptors in the aorta (de Cyon's nerve) and the carotid sinus (Hering's nerve). A decrease in pressure in these vascular sensory areas increases adrenaline secretion; hypertension diminishes it. Bilateral section of the aortic and carotid sinus nerves increases adrenaline secretion. The sensory mechanism of the arteries has the same effect on the adrenal medulla and its secretion as it has on sympathetic "tonus" (see "Physiologic significance of visceral innervation," Chap. 84).

Asphyxia provokes adrenaline discharge. The important factor is the fall in oxygen pressure; accumulation of CO_2 is of little importance. Inhibition of tissue oxidations by cyanide also provokes adrenaline secretion. Anoxia acts directly on the adrenaline secretory centers and reflexly by stimulating the vascular sensory areas. Denervation of the adrenals suppresses the adrenaline discharge of anoxia.[1] In the perfused gland, however, hypoxia provokes adrenaline discharge; therefore in this condition it acts directly on the chromaffin cells.[2]

Insulin hypoglycemia stimulates the adrenaline secretory centers, and adrenaline is discharged if the adrenal nerves are intact. The amount of adrenaline secreted is sufficient to produce hyperglycemia in another animal that receives the adrenal blood of the first[3] (Fig. 495). The increase in adrenaline secretion plays an important part in the recovery of the normal blood-sugar level following hypoglycemia. Adrenalectomized animals are hypersensitive to insulin.

Several drugs stimulate the adrenaline secretory centers. Among these are anesthetics (ether, urethane, etc.), strychnine, morphine, and picrotoxine. Drugs that have a nicotinic cholinergic activity (nicotine, acetylcholine, quaternary ammoniums, etc.) stimulate the chromaffin cells directly and provoke adrenaline discharge even after denervation of the gland. Other drugs increase adrenaline secretion by disturbing an organic equilibrium; substances

[1] HOUSSAY, B. A., and E. A. MOLINELLI, *Rev. Soc. argent. de biol.*, **6**, 402, 1925.

[2] BÜLBRING, E., *et al.*, *J. Physiol.*, **107**, 223, 1948.

[3] HOUSSAY, B. A., J. T. LEWIS, and E. A. MOLINELLI, *Rev. Asoc. méd. argent.* (Soc. Biol.), **37**, 486, 1934.

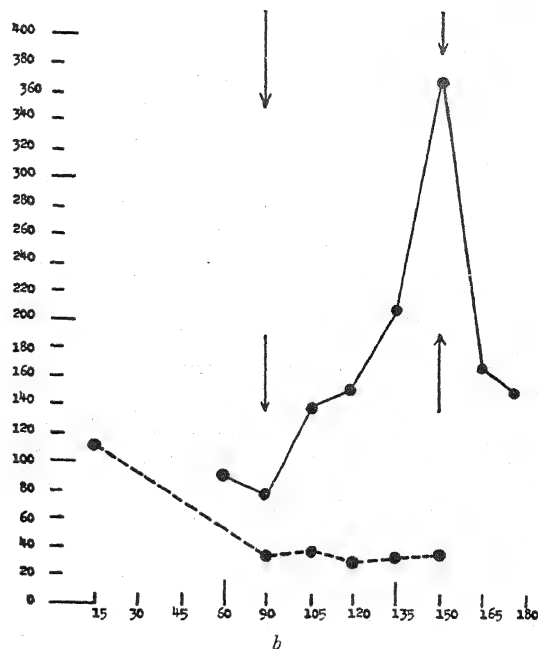
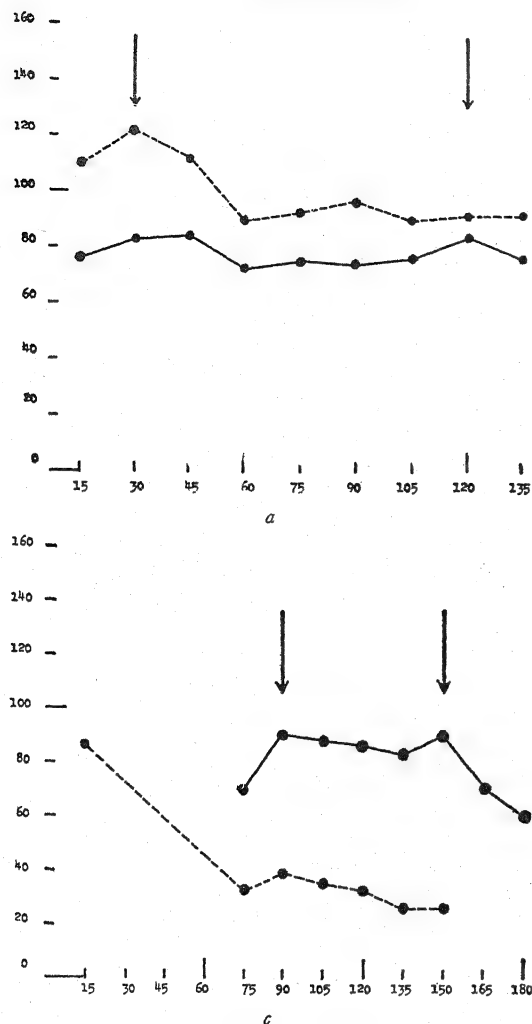



FIG. 495. Adrenojugular transfusion. Abscissa, time in minutes; ordinate, blood sugar in milligrams per cent. Solid line, receptor; broken line, donor. Between the arrows the transfusion takes place. *a*, normal controls; *b*, the donor was injected intravenously with 3 units insulin per kilogram 30 min. before the transfusion began; *c*, the donor's left adrenal was denervated and 3 units insulin per kilogram was injected intravenously 30 min. before the transfusion began. [*Houssay, B. A., J. T. Lewis, and E. A. Molinelli, Rev. Asoc. méd. argent.* (*Soc. Biol.*), *vol. 37, p. 486, 1924.*]

that produce hypotension (histamine) or hypoglycemia (insulin) act in this way.

Hypersecretion of adrenaline by adenomas of the adrenal medulla. Several cases of tumors made up of chromaffin cells (paragangliomas) have been reported. The patients have hypertensive crises provoked by exercise, emotional stress, and other conditions that normally increase adrenaline secretion. Removal of the tumor is immediately followed by a considerable fall in the blood pressure, which can be so marked and prolonged that it induces a condition of shock. While the paraganglioma is actively secreting adrenaline, the sympathetic tonus is kept low; the fall of arterial blood pressure on removal of the tumor is caused by the absence of sympathetic tonus. Usually after a few hours sympathetic tonus is recovered and the blood pressure returns to a normal level, but care must be taken to avoid the catastrophic hypotension that occurs after extirpating the tumor.[1]

Physiologic Significance of Adrenaline Secretion

In physiologic emergencies adrenaline is discharged into the blood. The increase in adrenaline secretion, produced through stimulation of the sympathetic, helps to reestablish the organic equilibriums that have been disturbed (homeostasis). Nevertheless adrenaline secretion is not of vital importance; it has been suppressed experimentally either by removing one adrenal and denervating the other or by extirpating the

[1] LERICHE, R., *Lyon chir.*, **31**, 355, 1934; BANER, J., and R. LERICHE, *Presse méd.*, **2**, 1384, 1932.

medulla,[1] the animals surviving in apparent good health. Nevertheless careful study of the animals deprived of adrenaline secretion shows that they are less adaptable to those circumstances which will provoke adrenaline secretion in the intact animal.

The importance of adrenaline secretion in conditions of emergency has been well demonstrated and is universally admitted. The same cannot be said about the secretion of adrenaline at rest, or in basal conditions, the existence of which Cannon has denied. Nevertheless the fact that adrenaline secretion is controlled in the same way as sympathetic tonus, by reflexes initiated in the vascular sensory areas, suggests that there is normally a basal secretion of adrenaline, apart from that artificially produced by experimental conditions. The physiologic significance of this basal secretion is so far unknown.

To summarize, adrenaline secretion is an auxiliary of the sympathetic, playing a prominent part in conditions of physiologic emergency. It is carried by the blood to the whole body and ensures the activation of all the effectors controlled by the sympathetic; it is thus an important factor in homeostasis.

[1] Houssay, B. A., and J. T. Lewis, *Rev. Asoc. méd argent.*, **34**, 1092, 1921.

BIBLIOGRAPHY

Cori, C. F., The Adrenal Medulla, *J. A. M. A.*, **116**, 2590, 1941.

Elliott, T. R., The Action of Adrenalin, *J. Physiol.*, **32**, 401, 1905.

Molinelli, E. A., "La Secreción de Adrenalina." Buenos Aires, 1926.

CHAPTER 86

Conditioned Reflexes

THE CLASSIC WORK of Pavlov and his associates on conditioned reflexes has given us new knowledge of the processes of association that take place in the cerebral cortex. It has also created a method for the objective study of cerebral activity which has been of great use not only in physiology but also in psychology. Pavlov did most of his work on dogs, but many other species including man have been studied. The relation of conditioned reflexes to another cortical function, *i.e.*, behavior, has been the object of much important work.

INBORN AND CONDITIONED REFLEXES

Reflexes are regular responses of the organism, which follow the application of certain stimuli in the same way as an effect follows the cause.

Reflexes are due to the transmission along definite nerve paths of an excitatory state from the receptor to the effector. They contribute to the maintenance of the internal bodily equilibrium and to the adaptation of the organism to its environment. They are related to instincts, which are much more complex phenomena but, according to Pavlov, do not differ fundamentally from reflexes.

Reflexes can be divided into two groups having typical characteristics. One group is made up of inborn, or unconditioned, reflexes and the other of acquired, or conditioned, reflexes.[1]

[1] All reflexes are "conditioned," inasmuch as certain definite conditions must be present for them to take place. The so-called "conditioned reflexes" could be more appropriately called "associated reflexes" or "reflexes acquired by association." The term introduced by Pavlov is now in common use and has a definite meaning; it should therefore be retained. When introducing new terms it is more important that the new meaning attached to them should be clearly defined than that they should be accurately descriptive.

Inborn, or unconditioned, reflexes are transmitted by heredity to all the individuals of the species, and they appear as soon as the nerve paths necessary for their performance are fully developed. They do not need the establishment of special conditions in the individual organism created by the previous application of stimuli. They follow certain definite preestablished paths and do not require the activity of the cortex. According to their object they may be considered as protective, postural, alimentary, sexual, etc. Some of them, such as the knee jerk, are simple two-neuron reflexes, others are much more complex, *e.g.*, the labyrinthine reflexes.

Acquired, or conditioned, reflexes do not appear spontaneously but are "learned" in the course of the individual's life. They are unstable, and they are easily lost either transitorily or permanently. They are established on another reflex, either inborn or previously conditioned. An unlimited number of stimuli acting on any of the receptors can provoke them, but they always follow nerve paths used by inborn reflexes. They appear after a formative process has been completed, in which the essential feature is the repeated application of the stimulus to be conditioned, associated with an efficient stimulus (Table 118 and Fig. 496).

Table 118. Characteristics of Inborn and Conditioned Reflexes

Inborn	*Conditioned*
Inherited by all the individuals of the species	Acquired by each individual
Stable	Unstable
Relatively few	Many
A special nerve path for each one	Nerve paths of other reflexes utilized
The cortex not indispensable	The cortex usually part of the reflex mechanism
No training needed	Training necessary

Pavlov demonstrated by the following experiment the difference between the two types of reflexes in the same individual: Five puppies were separated from their mother at birth and artificially fed on milk. After a few months they were shown meat for the first time. Although meat is the natural food of the species, they saw and smelled it without secreting saliva. Meat was then placed in their mouths, and saliva was secreted. The flow of saliva provoked by seeing or smelling meat is therefore a conditioned reflex acquired by the individual animal, and it does not exist until there has been a previous period of training. This reflex is built up on an inborn reflex, *i.e.*, salivary secretion provoked by meat placed in the mouth.

Fig. 496. Diagram of the path of a visual conditioned reflex. (*After Lovatt-Evans, modified.*)

Conditioned reflexes can be *positive*, *i.e.*, provoke excitation, or *negative*, *i.e.*, provoke inhibition.

CONDITIONS NECESSARY FOR THE FORMATION OF ACQUIRED REFLEXES

Positive and negative conditioned reflexes are produced by a special technique. They have been obtained in rabbits, sheep, dogs, cats, rats, monkeys, and man, and also in lower animals such as the tortoise, but the species most frequently used in this work has been the dog. Certain individual animals appear to be more intelligent and respond more readily than others; in some, positive reflexes are obtained more easily than negative reflexes, while the opposite is observed in others. A good physical condition is an important factor; reflexes are conditioned with difficulty, or not at all, in diseased animals. Psychic alertness is also necessary; distraction due to any cause prevents the establishment of these reflexes. Distraction is prevented by strict isolation. Pavlov placed his dogs in especially constructed soundproof chambers, which were also protected from any other kind of vibrations. The observer was placed in another room, from which he could see the animal through a periscope. The stimuli (food, light, sound, electrical stimulation of the skin, etc.) were applied and the effects (salivary flow, etc.) recorded by means of special devices (Fig. 497).

A great variety of stimuli, applied on any receptor, may provoke conditioned reflexes. Stimuli of great intensity, however, produce them with difficulty, and it is essential that they do not endanger the life of the animal. Among those that may be used are the following: auditory stimuli consisting in the vibrations of a tuning fork, giving a pure note, or complex sounds such as the blowing of a trumpet; visual stimuli varying in color, brightness, size, shape (*e.g.*, geometrical figures), duration, etc.; stimulation of the skin by tactile, thermal, and painful stimuli; and different odors to stimulate the sense of smell. In all these cases, if other necessary conditions are fulfilled, it is possible to build up a conditioned reflex.

The response can also be greatly varied, *e.g.*, muscular contractions, such as a knee jerk; a vasomotor reaction; dilatation or constriction of the pupil; secretion from one gland or another; etc. Pavlov frequently used salivary secretion as the response to be observed, because it can be easily collected from a fistula and accurately measured. The basis for this type of conditioned reflex is the inborn reflex of salivary secretion in response to food placed in the mouth. Defense reflexes, such as blinking when air is blown on the cornea, or the flexion reflex on application of a painful stimulus on the paw, are frequently used.

POSITIVE OR EXCITATORY CONDITIONED REFLEXES

There are many reflexes of this type, which vary according to the nature of the stimulus and the reflex on which they are based. A positive conditioned reflex is established by a procedure fundamentally the same as that followed in the conditioning of a sound to provoke salivary secretion: A parotid fistula is made in a dog, and after the wound has healed, the animal is placed in the soundproof chamber mentioned above, with a device adapted to the fistula so as to collect the saliva excreted. Meat is given

to the animal shortly after a bell has begun to ring. After this procedure has been repeated several times at suitable intervals, the sound of the bell alone will provoke salivary secretion, even if meat is not given. The sound of the bell, which previously did not provoke salivary secretion, has acquired the property of doing so by having been associated with an inborn salivary reflex. The auditory stimulus has become a conditioned stimulus (Table 119).

Table 119. Development of a Conditioned Reflex to a Sound of 637.5 c.p.s.

Number of times combination of sound and feeding had been performed	*Number of drops of saliva in 30 sec.*	*Latent period of reflex, sec.*
1	0	
9	18	15
15	30	4
31	65	2
41	69	1
51	64	2

Source: Anrep.

In the experiment referred to above, the originally neutral stimulus was applied at the same time as the stimulus that provoked the inborn reflex; this type of reflex is known as a *simultaneous conditioned reflex* (Fig. 498). If the originally neutral stimulus is applied some time before and maintained until the stimulus for the basic reflex is applied, the response is obtained after a latent period which lasts as long as the interval between the two stimuli and may be of several minutes duration. This type is known as a *delayed conditioned reflex.* If the conditioned stimulus precedes the stimulus for the basic reflex, but the latter is not applied until a short time after the former has ended, the reflex is called a *trace reflex.* The response is usually not so great as in the preceding types of reflexes. A stimulus can be conditioned using a previously conditioned reflex as a basic reflex, *e.g.*, the sound of a bell is conditioned to provoke salivary secretion, and when this reflex is well established, a colored light is shown to the animal before the bell rings. After this procedure is repeated a sufficient number of times, the colored light will become a conditioned stimulus and will provoke salivary secretion even if the bell does not ring. This is known as a conditioned reflex of the second order, and it can also be used as a basic reflex for the establishment of third-order conditioned reflexes, the latter

FIG. 497. Chamber for producing conditioned reflexes. The dog is in a soundproof compartment, completely isolated from the observer, who can see the animal through a periscope and controls the stimulating and recording apparatus without being seen by the animal. (*Pavlov, I. P., "Lectures on conditioned reflexes," trans. by W. Horsley Gantt, New York, 1928.*)

serving to build up conditioned reflexes of the fourth order, and so on. It is important that the conditioned stimulus should precede the stimulus for the basic reflex by an interval of approximately 10 sec. (Fig. 498).

e.g., if the sound of a tuning fork has been conditioned to provoke the secretion of 30 drops of saliva, and a tactile stimulus has been conditioned to provoke the same effect, *i.e.*, the secretion of 30 drops of saliva, simultaneous application of both stimuli will provoke the secretion of 60 drops of saliva.

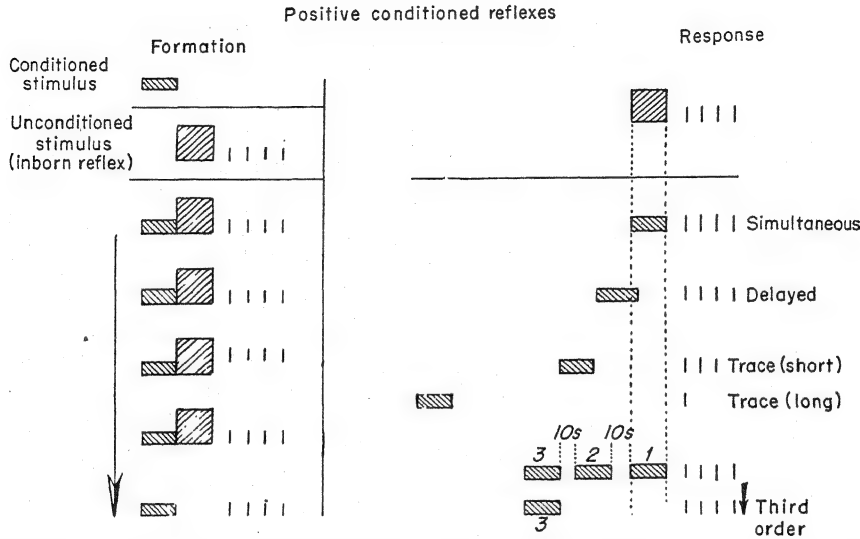



FIG. 498. Positive conditioned reflexes. On the left, the neutral stimulus at first provokes no response; the stimulus for the inborn reflex provokes a response. Both stimuli are applied together four times. The neutral stimulus is then applied alone, and a response is obtained; the animal has become conditioned to this stimulus. On the right, different types of conditioned reflexes obtained by varying the interval between the application of the neutral and the efficient stimulus.

The following are the essential conditions necessary for the establishment of a conditioned reflex: (*a*) the conditioned stimulus must be of at least threshold strength; (*b*) it must be applied repeatedly; (*c*) it must be applied at the same time as, or before, the stimulus for the basic reflex (if it is applied afterward it will not become conditioned); (*d*) it must be periodically reinforced, *i.e.*, accompanied by the stimulus for the basic reflex, otherwise it will rapidly lose its property of evoking the conditioned response (experimental extinction).

Conditioned reflexes have the following properties:

1. *Longer latent period; i.e.*, much longer than in inborn reflexes.
2. *Specificity*, *i.e.*, the response is evoked exclusively by the conditioned stimulus and not by other stimuli applied to the same or other receptors.
3. *Summation*, *i.e.*, two reflexes that use the same effector path in the same sense (excitation or inhibition) reinforce each other; *e.g.*, if the sound of a tuning fork has been conditioned to provoke the secretion of 30 drops of saliva, and a tactile stimulus has been conditioned to provoke the same effect, *i.e.*, the secretion of 30 drops of saliva, simultaneous application of both stimuli will provoke the secretion of 60 drops of saliva.
4. *Instability*, *i.e.*, they are easily lost, *e.g.*, when not reinforced.

Positive conditioned reflexes have been used in the study of the capacity of the sense organs of animals to respond. Thus it has been shown that dogs hear sounds of a much higher pitch (greater frequency of vibration) than can be heard by the human ear. Combining the conditioned-reflex method with the production of lesions limited to a definite part of the receptor (*e.g.*, the cochlea) or the nerve centers (*e.g.*, the cerebral cortex), it has been possible to localize the site in the receptor where certain stimuli produce their effects and the nerve paths and centers of the different sense organs.

NEGATIVE OR INHIBITORY CONDITIONED REFLEXES

Well-established conditioned reflexes can be extinguished (*i.e.*, suppressed) by processes of external or internal inhibition.

External inhibition. If during or shortly before the application of a conditioned stimulus another stimulus of any kind is applied, which provokes another reflex or simply distracts the attention of the animal, the response to the conditioned stimulus is suppressed or considerably diminished. This is known as external inhibition, or interference. It is rapidly established but acts only temporarily, and if the "extrastimulus" is repeated it soon loses its inhibitory effect (Fig. 499). External inhibition may be the cause of fluctuations in the response to conditioned stimuli. Pavlov discovered it when he tried to make a public demonstration of conditioned reflexes; the animals did not respond, and the demonstration could not be made. This led him to realize the importance of isolating the animals strictly from other stimuli besides those which are being studied.

Internal inhibition. This type of inhibition differs from external inhibition because it is necessary to apply the stimulus repeatedly before it becomes efficacious; it differs also in the slowness with which it is established. Four types of internal inhibition are distinguished, according to the method by which they are established: (*a*) extinction; (*b*) conditioned inhibition; (*c*) inhibition in delayed and trace reflexes; (*d*) differential inhibition.

Extinction. Positive conditioned reflexes that are not reinforced are soon lost. For example, if an animal is conditioned to secrete saliva on hearing the sound of a tuning fork by applying this stimulus simultaneously with feeding or before it, and if when the conditioned reflex is well established the animal is made to hear the tuning fork repeatedly without being given food, saliva is secreted in progressively decreasing quantities until the sound of the tuning fork no longer provokes salivary secretion (Fig. 500). The extinction of the reflex is due not to fatigue but to inhibition. This can be demonstrated by applying an extrastimulus that will "dis-inhibit" the animal, *i.e.*, inhibit the inhibitory state that prevents salivary secretion; a copious flow of saliva will then be observed.

Extinction occurs more easily in recently established conditioned reflexes than in long-standing ones. The strength of inhibition can be measured by the time needed to reestablish the reflex. Extinction has an "after-effect" which increases the difficulty of obtaining a positive conditioned reflex.

Conditioned inhibition. If a strong extrastimulus is applied simultaneously with a conditioned stimulus without reinforcing the latter, after this procedure has been repeated several times a negative conditioned reflex will be established (Fig. 500), *i.e.*, the application of the extrastimulus will inhibit salivary secretion, although the application of the originally conditioned stimulus alone still provokes the secretion of saliva. The central inhibitory state thus created has a fairly prolonged after-effect, during which it is difficult to obtain a positive conditioned reflex. The inhibitory state can be inhibited by an extrastimulus, and its suppression provokes the reappearance of the underlying positive reflex.

FIG. 499. External inhibition. Percentage inhibition of a positive conditioned response after repeated application of an extrastimulus.

Conditioned inhibition is similar to extinction in that to provoke both of them the conditioned stimulus must not be reinforced. Conditioned inhibition and external inhibition are similar in that both require the addition of an extrastimulus. In external inhibition, however, the extrastimulus exerts its inhibitory effect immediately and then ceases to be efficacious, whereas in conditioned inhibition the extrastimulus

acquires gradually inhibitory properties owing to the absence of reinforcement.

Inhibition in delayed and trace reflexes. During the interval between the two stimuli in delayed and in trace reflexes, a state of conditioned inhibition is developed. As the interval is increased in trace reflexes the positive response diminishes until it is completely suppressed. If the interval is very long, *e.g.*, 30 min., the inhibitory state extends to other reflexes and the animal falls asleep. An extrastimulus applied in this condition suppresses inhibition and provokes the appearance of a positive response. This type of inhibition is obtained more easily with cutaneous and visual stimuli than with auditory ones.

Differential inhibition. Discrimination. There are two stages in the building up of a positive conditioned reflex. During the first stage a response is provoked not only by the specific stimulus but also by similar stimuli applied to the same receptor. In the second stage there is discrimination; only the specific stimulus provokes a positive response; responses to similar stimuli are inhibited. This process can be accelerated by applying the specific stimulus and reinforcing it, together with accessory stimuli without reinforcement, so that the latter undergo experimental extinction. By means of this method it is possible to determine objectively an animal's capacity to distinguish between two stimuli, *i.e.*, the analytical capacity of the central nervous system, according to Pavlov. The different mechanisms responsible for discrimination, *i.e.*, the receptors, nerve paths, and centers, have been called by Pavlov the "analyzers."

The auditory analyzer discriminates differences in pitch, intensity, quality, and direction with greater precision in the dog than in man. The visual analyzer of the dog discriminates differences in degree of luminosity, in the shape and position of objects, and in their movements, but differences in color are not perceived if there is no difference in luminosity. The cutaneous analyzer can distinguish degrees of smoothness or roughness of a surface, differences in temperature of about ½ to 1°C., the place where the stimulus is applied, the strength of pressure, and the duration of the stimulus.

In the process of establishing discrimination, internal inhibition plays an important part; in some cases inhibition spreads throughout the whole cortex, and the animals fall asleep. If the problem of discrimination exceeds the capacity of the animal, it is not only unable to establish final discrimination, but the whole structure of its nervous function is disturbed. For example, if the animal is made to distinguish a circle from a long oval, discrimination is perfect. If the oval is made progressively broader, so that finally there is little difference between the oval and a circle, not only does the animal not distinguish them but the whole pattern of its conditioned reflexes is upset (*e.g.*, inhibitory stimuli become positive and vice versa) and it is impossible to establish new conditioned reflexes. Moreover the animal becomes excitable and bad-tempered. Its behavior is similar to that of neurotic patients; hence the name of "experimental neurosis" given by Pavlov to this condition. Chimpanzees become angered, or they may remain indifferent and refuse to pay attention. Rats, rabbits, pigs, and sheep also show this phenomenon, but the behavior varies and is typical for each species. Disturbances in pulse rate, breathing, metabolism, urinary

FIG. 500. Internal inhibition. Left, experimental extinction is due to repeated application of the conditioned stimulus (*CS*) without reinforcement by the unconditioned (*UCS*) or efficient stimulus; at the fifth application there is no response (*R*). Center, conditioned inhibition is obtained by the application of an extrastimulus together with the positively conditioned stimulus but without reinforcement. Right, if the delay is gradually lengthened, the response will be inhibited.

secretion, and sexual functions have been observed, lasting up to 24 hr. after the stimulus was applied. Double frontal lobectomy causes fundamental changes in the behavior of these animals, which are no longer angered or disturbed when their discriminatory powers are exceeded.

General characteristics of inhibition. Inhibitory processes prevent reflex responses that have no useful purpose. Repeated application of the conditioned stimulus in experimental extinction and the application of a second stimulus in conditioned inhibition imply circumstances in which the positive response is no longer useful. Inhibition in delayed and trace reflexes retards the response until it can be useful. Differential inhibition in the process of discrimination also signifies the suppression of responses that are not adapted to the stimulus. Thus inhibition is an important mechanism in the prevention of wasteful expenditure of energy and matter.

Internal inhibition is not merely fatigue produced by repetition of the stimulus; it is an active process that takes place in the nerve centers. Proof of this statement is given by the following facts:

1. Inhibition can be suppressed by an extra-stimulus, and the positive response will appear; *e.g.*, a strange noise produced during the latent period of a trace reflex that produces salivary secretion provokes the immediate appearance of the response.
2. Negative (inhibitory) conditioned reflexes have an after-effect, *i.e.*, they are followed by a period during which it is very difficult to obtain a positive response to other conditioned reflexes. The inhibitory state produced by the negative reflex extends first to other reflexes in the same analyzer and then to other analyzers, gradually spreading throughout the whole cortex and to subcortical centers, eventually producing a condition similar to, if not identical with, sleep. The wave of inhibition then recedes and is reduced to the analyzer corresponding to the inhibited reflex (not to the analyzer that received the inhibiting stimulus).

NERVE CENTERS OF CONDITIONED REFLEXES

The establishment of a conditioned reflex requires the integrity of the whole analyzer, *i.e.*, not only of the receptor, of the peripheral paths that conduct the nerve impulses to the subcortical centers, of these centers, and of the central paths to the cortex, but also of the cortical centers. Pavlov maintained that the cortex was essential to the production of conditioned reflexes, because extirpation of the cortex suppresses all the conditioned reflexes acquired by the animal and also the possibility of acquiring new ones, although inborn reflexes remain. Later experiments carefully carried out by Zeljony and by Culler and Mettler have shown that it is possible to build up a few simple auditory, visual, and tactile conditioned reflexes in totally decorticate dogs and cats. Experimental extinction and discrimination, *i.e.*, inhibition, are characteristic of conditioned reflexes in these animals just as in animals with an intact cortex. It is therefore evident that in certain cases subcortical centers can replace the cortex in the associative function necessary for the building up of conditioned reflexes.

Partial extirpation of the cortex, *e.g.*, of both visual areas or both auditory areas, is followed by marked disturbances in the conditioned reflexes corresponding to the damaged analyzers. Immediately after the operation, all the conditioned reflexes of the animal are upset, especially those involving inhibitory processes, but a few days later they are reestablished, with the exception of those corresponding to the area extirpated. A few simple reflexes corresponding to these areas are nevertheless recovered. Thus Marquis and Hilgard have observed that dogs and monkeys in which the visual areas of the cortex have been removed can distinguish differences in the brightness of light, and that those which have undergone extirpation of the auditory areas can distinguish sounds of different pitch. Moreover a few new reflexes can be built up in these decorticate animals.

The site of association of the efficient with the neutral stimulus and the nature of the process by which they become associated are not known. This association does not take place at the level of the motor neuron, because it is not possible to establish a conditioned reflex on the basis of stimulation of this neuron (Loucks).

The behavior of animals confronted with a problem such as opening a box or finding their way in a maze has been of great use in the study of cortical function.

The work of Lashley,[1] Marquis, Hilgard, and others in rats, dogs, and chimpanzees has shown that extirpation of a limited area of the cortex does not produce any serious deficiency in performance because the functions of the area removed are taken up by other cortical and subcortical centers. Strict specialization of structures seems to be less important than the mass of functional tissue. Even in areas where there is a high degree of specialization (*e.g.*, the motor and visual areas), the subordinate parts can take on the functions of the whole. On the other hand, functional deficiency increases as larger areas are removed and the functioning mass of cortex diminishes. Conditioned reflexes and behavior are closely connected, and the knowledge acquired by work done in one field is in agreement with that in the other and throws light on it.

In summary, conditioned reflexes are built up by a central mechanism of association which takes place in cortical and subcortical centers.

THE EFFECT OF DRUGS ON CONDITIONED REFLEXES

Caffeine (25 to 50 mg. in dogs) increases all conditioned reflexes and diminishes inhibition. The effect begins 20 to 30 min. after the drug has been administered and lasts 24 hr. Strychnine has a similar action.

Bromides (1.5 gm.) have no effect on conditioned reflexes, but they increase inhibitory processes. Their action lasts for 2 or 3 days, owing to the slow rate at which they are eliminated.

SUMMARY

Conditioned reflexes are the result of the experience of the individual. They are of great importance for adaptation because conditioned stimuli are signs of forthcoming events for which the organism can anticipate an adequate response. Positive reflexes widen the scope of inborn reflexes by continuously adding responses to new and unaccustomed stimuli. Inhibitory reflexes are of great importance for accuracy and fineness in adaptation and for the conservation of energy and matter. A dog deprived of the cerebral cortex has very limited possibilities with respect to acquiring conditioned reflexes; Pavlov considered such an animal as a "great invalid." The establishment of conditioned reflexes has provided an objective method for the study of (*a*) the localization of sensory nerve centers and paths; (*b*) cortical and subcortical mechanisms of association; (*c*) the capacity and acuteness of the different analyzers in many animal species. It has also contributed facts of importance for the understanding of sleep and neurosis.

[1] LASHLEY, K., Mass Action in Cerebral Function, *Science*, **73**, 245, 1931.

BIBLIOGRAPHY

ALVAREZ-BUYLLA, R., Descripción de una Cámara para el Estudio de Reflejos Condicionados, *Ciencia*, **10**, 291, 1950.

HILGARD, E. R., and D. G. MARQUIS, "Conditioning and Learning," Appleton-Century-Crofts, New York, 1940.

HILGARD, E. R., and C. P. STONE, Physiological Psychology, *Ann. Rev. Physiol.*, **1**, 471, 1939.

HORSLEY GANTT, W., "Experimental Basis for Neurotic Behavior," Psychosomatic Medical Monographs, Vol. 3, Nos. 3 and 4, 1944.

JACOBSEN, C. F., The Effects of Extirpation on Higher Brain Processes, *Physiol. Rev.*, **19**, 303, 1939.

LASHLEY, K., *Physiol. Rev.*, **13**, 1, 1933.

LIDDELL, H. S., The Nervous System as a Whole: The Conditioned Reflex, in Fulton's "Physiology of the Nervous System," 3d ed., Oxford, New York, 1949.

LIDDELL, H. S., Physiological Psychology, *Ann. Rev. Physiol.*, **3**, 487, 1941.

LOVATT-EVANS, C., "Recent Advances in Physiology," 4th ed., J. & A. Churchill, London, 1930.

PAVLOV, I. P., "Conditioned Reflexes," Oxford, New York, 1927.

CHAPTER 87

The Electroencephalogram

THE CEREBRAL CORTEX continuously shows electric potential waves which are the resultant of the synchronized activity of a great number of neurons. They are due to a tendency to synchronized rhythmicity observed in central groups of neurons in vertebrates and invertebrates. Electric potential waves have also been picked up from the cerebellum, thalamus, diencephalon, geniculate bodies, and olfactory lobe and in the neural ganglia of invertebrates.

Frequency is the most typical characteristic of the electric activity of the brain and the one that has been most easily correlated with cortical function (Gibbs). It varies from 1 to 50, with voltages from 5 to 500 μv (*i.e.*, much lower voltages than those of the electric activity of the heart). Several rhythms of different frequencies blend into the complex potential of the electroencephalogram, but it is possible to separate and analyze different rhythms by means of special devices (Walter *et al.*). The EEG changes in different physiologic, psychic, and pathologic circumstances (Figs. 502, 503, and 504).

These potentials are evidence of the constant activity of the nerve centers, which are never completely at rest. In some cases fluctuations in the potentials coincide with the automatic rhythmic activity of the center. The waves from the cortex persist after the brain stem has been cut, *i.e.*, after many of the afferent impulses to the cortex have been eliminated.[1]

The electrical activity of the cortex was registered for the first time in 1874 and was repeatedly observed after that date,[2] but its registration and study in man begin with the observations of Hans Berger.[1] The record of the electrical activity of the brain is known as the electroencephalogram (EEG). It is obtained by means of an electroencephalograph, and the study of this electrical activity is known as electroencephalography. The record obtained by placing the electrodes directly on the cerebral cortex after the cranium has been opened in the course of an operation (the cortical electrogram) does not differ significantly from the record obtained by placing the electrodes on the skin of the head.

[1] BREMER, F., *Compt. rend. Soc. de biol.*, **118,** 1235 and 1241, 1935; "L'activité électrique de l'écorce cérébrale," Hermann & Cie, Paris, 1938; *J. belge de neurol. et de psychiat.*, **47,** No. 9, 1947.

[2] Caton, 1874; Fleisch von Marxow, 1883; Gotch and Horsley, 1892; Berger, 1902; Neminski, 1925; Berger, 1924–1929; Adrian, 1934; Davis, Gibbs, Jasper, Lennox, 1935; Walter, 1936, etc.

Nonpolarizable electrodes are not necessary. The electrodes are small metal plates applied on the skin of the head, which has been previously wetted with saline; they are kept in position by means of collodion. The electrodes are connected with a cathode-ray oscillograph or other recording device, by means of amplifiers. The inscription is usually made by ink-writing recorders. Bipolar leads, with both electrodes placed over the cortex, or monopolar leads, with one electrode on an ear and the other over the cortex, are employed. In both cases the electrical activity of the cortical areas near the electrodes is recorded; that of more distant areas exerts only a weak influence or none on the record. Two to four electrodes are placed on each side of the head over the frontal, parietal, temporal, and occipital areas. The electric activity of the base of the brain is recorded by means of electrodes placed in the nasopharynx or on the tympanic membrane.

THE NORMAL ELECTRO-ENCEPHALOGRAM

The cortical electric pulse appears as a more or less sinusoidal oscillation, the frequency of which varies from 1 to 50 c.p.s. (Fig. 501), or an

[1] BERGER, H., *Arch. f. Psychiat.*, **87,** 527, 1929; **101,** 454, 1933.

even wider range if a good recording instrument is used. In the adult the voltage is usually from 10 to 50 μv, never more than 200 μv. Since the early days of encephalography, the waves have been classified according to their frequency in

FIG. 501. Frequency of cortical potential waves in different normal and pathologic conditions. (*Gibbs, F. A., and E. L. Gibbs, "Atlas of Encephalography," privately printed by the authors, Cambridge, Mass.*)

three groups named with the Greek letters α, β, and δ; it is better to consider them as a continuous spectrum of frequencies (Fig. 501). Several authors do not use the Greek-letter notation (alpha, beta, gamma, delta, theta) but give the frequency per second. Six wave bands are usually distinguished (Table 120).

In the normal EEG an α rhythm is predominant, with waves of an average frequency of 10 (8 to 13 per second) and a voltage of 20 to 50 μv (Table 120, Fig. 502). This is known as

Table 120. Wave Bands in the EEG

	δ		α		β	*Fast*
Frequency per sec.....	0.5–3	4–7	8–13	14–17	18–60	Over 30
Amplitude μv	20–200	...	20–50	.....	5–10	

Berger's rhythm; it is the usual one found in healthy resting subjects with their eyes shut. There is a more or less marked modulation of the α waves, which increase and decrease periodically. When this is marked, spindle-shaped groups of increasing α waves, followed by decreasing ones, are formed.

The β waves are more frequent (18 to 60 per second) and of lower voltage (5 to 10 μv). The δ waves are slow; there are 0.5 to 5 per second, and they are of relatively high voltage (20 to 200 μv). Normal waves are sinusoidal, but in abnormal cases large waves, with a round or flat summit, or spikes are observed (Fig. 504). The EEG does not lose its normal features after the three outer layers of the cortex have been destroyed; but if the deeper layers, *i.e.*, the large pyramidal cells, are destroyed, the normal rhythm disappears.

The EEG varies in different circumstances. The most important only will be mentioned here.

Age. The EEG of the newborn child shows slow irregular waves (0.5 to 2 c.p.s.) on which fast waves of low voltage are superimposed. As the child grows the waves become more regular and faster. At the age of nine or ten years the occipital record shows an α rhythm (10 per second), but up to nineteen years of age the full adult pattern may not have been attained.

Cortical areas. Kornmuller maintained that in the rabbit each cytoarchitectonic area of the cortex had a typical EEG. Such an extreme opinion is not shared by many observers, but in man regional differences have been reported. Thus records from the occipital area have a predominant α rhythm;[1] β waves (15 to 30 per second) are predominant in the frontal and parietal areas. An α rhythm is not always the predominant one in the occipital area, and it may be found in other areas.

Individual differences. The EEG of one individual differs from that of others, but the difference is not such that it is always possible to distinguish the EEG of one person from that of another with the same type of EEG. The individual pattern of the EEG remains stable. The electroencephalograms of identical twins are as similar as two records taken from the same individual at different times.

Levels of cortical activity. The EEG changes with the level of cortical activity. A subject physically and mentally resting with eyes closed has an α rhythm, *i.e.*, relatively ample waves (20 to 50 μv) of low frequency (8 to 13 per second). This is the typical EEG when *awake and resting* (Fig. 502). If the eyes are opened and

[1] ADRIAN, E. D., and B. H. C. MATTHEWS, *Brain*, 57, 355, 1934.

fixed on an object, or if in the dark an effort to see is made, or if the mind is concentrated on a problem, the α rhythm is replaced by small waves of higher frequency (20 to 50 per second). This is the EEG typical of general *alertness* (Fig. 502). Blocking of the α rhythm is a relatively slow process; it takes approximately half a second (0.45 sec.) to be completed. If no special stimuli are applied and the subject's attention wanders, the α rhythm reappears even if the eyes are kept open in a normally lit room. In a *somnolent* state the α rhythm gives way to a slow rhythm of 4 to 6 waves per second. Bursts of waves of progressively increasing and decreasing amplitude and 14 to 16 per second frequency suddenly appear; these are the so-called "sleep spindles." Occasionally large slow waves are seen (Fig. 503). During *sleep* only high-voltage slow waves (0.5 to 3 per second) are observed. These typical waves of normal sleep should not be mistaken for the large abnormal waves seen in subjects with a tendency to convulsions (Gibbs). *Hypnotics* and *anesthetics* have the same effect as sleep on the EEG.

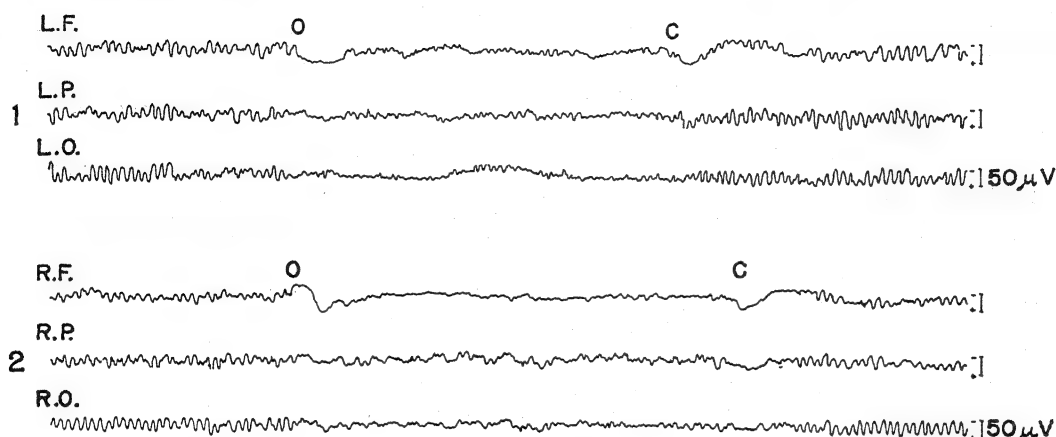



Fig. 502. Electroencephalograms of adults with eyes open (*O*) and closed (*C*). (*Gibbs, F. A., and E. L. Gibbs, "Atlas of Encephalography," privately printed by the authors, Cambridge, Mass.*)

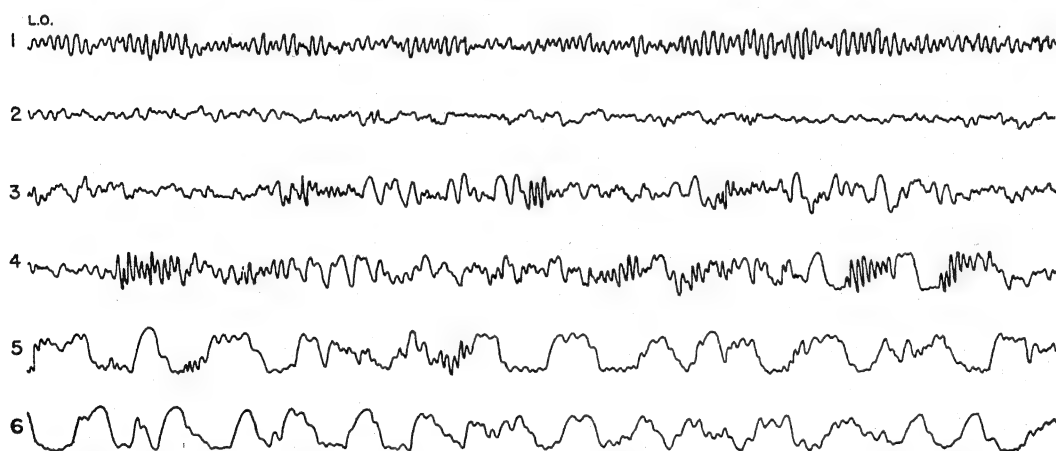



Fig. 503. EEG during sleep. 1, subject is awake (11:30 P.M.); 2, falling asleep (11:45 P.M.); 3, light sleep (12:00 M.); 4, moderately deep sleep (12:30 A.M.); 5, deep sleep (1:00 A.M.); 6, very deep sleep (2:00 A.M.). (*Gibbs, F. A., and E. L. Gibbs, "Atlas of Encephalography," privately printed by the authors, Cambridge, Mass.*)

When a sleeping subject awakens, the opposite changes take place. The slow rhythm gives place to another of higher frequency and lower voltage, the α rhythm, or that corresponding to alertness, according to the level of cortical activity. This *awakening reaction* does not begin in

any special part of the brain, not even in the cortical centers of the sensory area the stimulation of which has provoked it. Thus the EEG of awakening has been obtained in cats by stimulation of the vestibular apparatus after the respective cortical centers have been destroyed. On awakening, as in falling asleep, the whole cortex responds at the same time.

The large waves of the EEG are attributed to synchronization of the electrical activity of many neurons. Depression of cortical activity is accompanied by "recruitment" of neurons which enter into phase, and their electrical activity is synchronized. Activity, on the contrary, "disbands" the neurons and their electrical activity is desynchronized.

Significance of the waves. The spontaneous electric activity of the cortex has been attributed to (*a*) automatic, rhythmic activity of the neurons, especially of the large deep pyramidal cells; (*b*) rhythmic stimulation of the cortex by volleys of impulses from a diencephalic pacemaker; (*c*) impulses passing through the cortex in reverberatory action of thalamocortico-thalamic circuits.

The activity of subcortical centers modifies the EEG. This can be demonstrated by (*a*) simultaneous registration of cortical and subcortical phenomena; (*b*) section of paths between the cortex and subcortical nuclei; (*c*) observing the effects of stimulation of subcortical nuclei on the EEG.

A thalamocortical component has been demonstrated (Morison and Dempsey). There is an ascending activating system in the reticular formation, from the medulla to the diencephalon (Magoun *et al*). Sensory impulses also exert a certain influence (Adrian and Matthews).

Bremer maintains that the EEG is due to the automatic, rhythmic, synchronized activity of cortical neurons, sustained by continuous afferent impulses coming mainly from the thalamus.

THE EEG IN ABNORMAL CONDITIONS

"Normal abnormals." In 10 per cent of apparently normal subjects, the EEG shows changes similar to those seen in 90 per cent of epileptics. Some of these subjects are "difficult" children; others have disturbances in personality. According to some authorities, they are asymptomatic epileptics (Gibbs). The percentage of cortical dysrhythmia is significantly higher in the families of epileptics than in the general population.

Epilepsy. The EEG and other methods of exploration have shown that there is persistent cerebral dysrhythmia in epileptics.[1] The rhythm is either abnormally slow or abnormally fast, and usually fast and slow rhythms alternate. Regulation of wave frequency, *i.e.*, of the energy discharge, is disturbed, and there is a tendency to abnormally extensive and violent discharges. The three main types of epilepsy show distinctive patterns in the EEG.

Petit mal, with transitory disturbance or loss of consciousness and only slight spasms in the facial muscles, is characterized by an EEG with slow rhythm (3 per second) and waves of high voltage, in which usually a slow wave of rounded summit alternates with a spike potential (Fig. 504). When this rhythm is not in evidence it can be provoked by hyperventilation or hypoglycemia, and when present it has a tendency to disappear when the subject breathes 3 to 4 per cent CO_2. Stimulation of a small area of the thalamus provokes bilateral cortical discharges of waves and spikes.

Grand mal is the typical epileptic seizure with tonic followed by clonic convulsions, loss of consciousness, dilatation of the pupil, frothing at the mouth, relaxation of sphincters with involuntary emission of urine, etc. Slow waves appear before the seizure, and when the attack begins there is a sudden decrease in the amplitude of the waves. During the stage of tonic convulsions, voltage and frequency increase progressively; the frequency may be 15 to 40 per second. During the clonic convulsions slow waves predominate, and immediately after the seizure the record is flat without waves, or only a few slow waves are present. After a few minutes the subject's usual rhythm is reestablished (Fig. 504). Similar changes in the EEG have been observed in several types of experimental epilepsy.[2]

Psychomotor epilepsy is the type in which the seizure takes on a stereotyped behavior pattern, *e.g.*, an emotional outburst, an attack of rage or fear, flight, amnesia, somnambulism, etc. The EEG (Fig. 504) shows series of slow waves (2 to 4 per second) of high voltage with flat or serrated

[1] Gibbs, F. A., E. L. Gibbs, and W. G. Lennox, *Arch. Neurol. & Psychiat.*, **39**, 298, 1938.

[2] De Finis, M. L., and J. B. Odoriz, *Rev. Soc. argent. de biol.*, **19**, 217, 1943.

summits, on which fast (14 per second) waves are superimposed; finally only the slow waves are seen. Slow waves are observed 30 times more frequently in epileptics than in normal subjects.

The EEG gives important data for the diagnosis, localization, prognosis, and management of epileptics. In "difficult children" and in apparently normal members of epileptic families, it shows dysrhythmias and cortical disturbances that are not revealed by any other method.

In cases of epilepsy the area of the cortex in which the seizure begins can be localized by means of the EEG. The electrical disturbance (usually slow waves) starts in a definite area and may either spread to other areas or remain localized, being in the latter case the only manifestation of the attack. In certain cases, by taking simultaneous records from several areas, it is possible not only to determine the site of the focal discharge but also to follow the path along which the disturbance spreads.

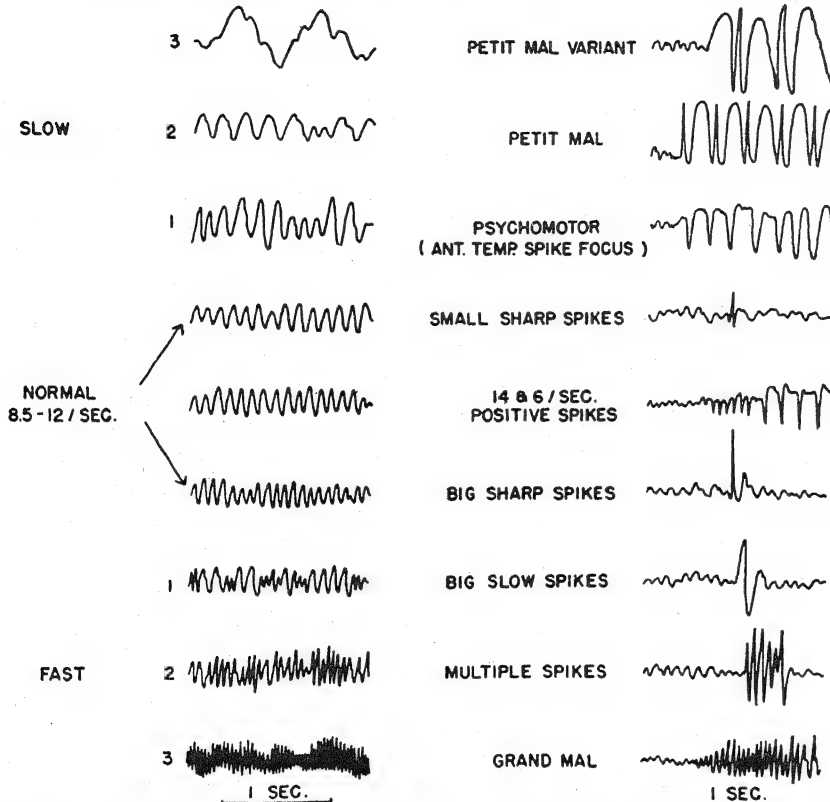



FIG. 504. Types of normal and abnormal electroencephalogram. Left-hand column, nonparoxysmal patterns: above, those that are abnormally slow for adults; below, those that are abnormally fast; in the middle, patterns that fall in the 8½- to 12-per-sec. range, which is normal for adults. The scale of abnormality reads from 1 to 3, with 1 being slightly abnormal, 2, very abnormal, and 3, exceedingly abnormal. This scale, however, must be adjusted for age. Right-hand column, paroxysmal patterns, given descriptive names or named after the clinical type of seizure with which they are most commonly associated. The pattern referred to as "psychomotor" appears as a generalized disorder over the entire cortex during a clinical seizure of the psychomotor type. In seizure-free intervals, particularly during sleep, the psychomotor disorder appears as a focus of negative-spike seizure potentials in the anterior temporal region. The 14- and 6-per sec. positive spikes are in reality two patterns, which can appear independently as 16- or 6-per-sec. positive spikes. (*F. A. Gibbs, Bull. New York Acad. Med., p. 771, 1949.*)

Localization of brain lesions. Focal lesions in the cortex, such as those due to a brain tumor, subdural hematoma, abscesses, meningitis, encephalitis, etc., can be localized by means of the EEG. The method is especially valuable for showing the existence of cortical damage in cases of cranial trauma, Sydenham's chorea, encephalitis lethargica, etc. In recent years attempts have been made to localize subcortical

lesions by means of the EEG. In 90 per cent of cases of brain tumor it is possible to make an accurate diagnosis of the location of the tumor (Gibbs). Records are made from different parts of the cortex, and those of symmetric points on the hemispheres are compared. When there is focal damage the records taken from the damaged area or from neighboring areas show certain changes, usually slow waves and spikes.

Electroencephalography is also being applied to the study of psychiatric problems and in legal medicine.

Chemical factors. Many physical and chemical agents that have effects on metabolic processes provoke changes in the EEG. Thus anoxia, alkalosis, and hypoglycemia diminish the frequency of the waves and may completely abolish them. Acidosis, excess CO_2, and light ether increase the frequency of the waves. The effect of drugs on the EEG has also been studied. Thus changes similar to those seen in epilepsy are observed when convulsive doses of cortical stimulants are given.

BIBLIOGRAPHY

ADRIAN, E. D., *Proc. Roy. Soc., London, s.B.*, **126**, 433, 1939.

BRAZIER, M. A., "The Electrical Activity of the Nervous System," Sir Isaac Pitman & Sons, Ltd., London, 1951.

BREMER, E., *Electroencephalog. & Clin. Neurophysiol.*, **1**, 177, 1949.

BREMER, E., *J. belge de neurol. et de psychiat.*, **47**, 1, 1947.

DAVIS, H., Electrical Phenomena of the Brain and Spinal Cord, *Ann. Rev. Physiol.*, **3**, 377, 1939.

GIBBS, F. A., and E. L. GIBBS, "Atlas of Encephalography," 2d ed., Addison-Wesley Press, Cambridge, Mass., 1950–1951.

GIBBS, F. A. and E. L. GIBBS, *Bull. New York Acad. Med.*, **25**, 765, 1949.

JASPER, H., Electrical Activity of the Brain, *Ann. Rev. Physiol.*, **3**, 377, 1941.

JASPER, H. Deuxième Congrés internat. d'électroencephalographie, Paris, 1949.

LENNOX, M. A., and W. G. LENNOX, Electrical Activity of the Brain, *Ann. Rev. Physiol.*, **9**, 507, 1947.

SCHWAB, R. S., "Electroencephalography in Clinical Practice," Saunders, Philadelphia, 1951.

TOMAN, J. E. P., and J. P. DAVIS, The Effects of Drugs on the Electrical Activity of the Brain, *Pharmacol. Rev.*, **1**, 425, 1949.

WALTER, W. G., *et al.*, "Electroencephalography. A Symposium on its Various Aspects," Macmillan, New York, 1950.

WALTER, W. G., and V. I. WALTER, The Electrical Activity of the Brain, *Ann. Rev. Physiol.*, **11**, 199, 1949.

CHAPTER 88

Sleep

SLEEP IS A PERIODIC and reversible physiologic state, which is necessary for the repair of the organism. During sleep consciousness is depressed, the threshold of sensory perception is considerably increased, and spontaneous movements and muscle tone are reduced to a minimum, while metabolism and visceral functions are not markedly diminished. The outstanding features of sleep are (*a*) absence of response to ordinary changes in the environment because the sensory threshold is raised; (*b*) unconsciousness, which can be dispelled by adequate stimulation; (*c*) muscular relaxation; (*d*) suppression of higher cortical activity; (*e*) total or partial suppression of normal mental activity, with the occurrence of dreams during light sleep. Periodic depression of activity or transient stages of resting may be observed in lower forms of life, including protozoa and plants. Such periods have been compared to the sleep periods of the higher vertebrates, including man. In fishes, for instance, characteristic "sleep positions" have been described.

The onset of sleep. The passage from the waking state to sleep takes place gradually. Drowsiness begins by an increasingly strong sensation of fatigue, indifference to environmental stimuli, and depression of the sensory activity. Motor responses become slow and inaccurate, and muscle tone diminishes. The eyelids drop over the eyes, the head nods fitfully, the jaw hangs down, the trunk becomes flexed, and the arms are pendant. Relaxation of the soft palate causes snoring. Immediately before sleep distorted visual and auditory sensations may be experienced.

The onset of sleep is favored by the suppression of stimuli and by facilitating muscular relaxation. Thus silence, darkness, an adequate temperature, a recumbent or sitting position, and mental composure are conducive to sleep. On the other hand, motor activity, strong stimuli, and mental worry or excitement retard the onset of sleep.

Duration and depth of sleep. On the average, man passes about one-third of his life asleep. The daily requirement of sleep varies with age and in different individuals. The young infant sleeps 20 hr. daily. Between the first and seventh year, sleep duration is reduced gradually from 14 to 10 hr. per day. On the average, adults require 7 to 9 hr. sleep in 24 hr. and old persons only 5 to 6 hr. In most adults the depth of sleep increases to a maximum during the first hour and then gradually diminishes until the time of awakening; in a large number of individuals there sets in a second period of deep sleep before waking. Sleep during the daytime is usually lighter than at night. The great majority of people sleep at night and are awake during the day, but the cycle can be reversed, as in the case of people who work at night, and in warm climates a short period of sleep at midday is customary.

PHYSIOLOGIC CHANGES DURING SLEEP

Sleep is accompanied by marked changes in somatic functions and by some changes in visceral activity.

Muscular activity. All the muscles are relaxed, except the flexors of the hand in children. There is some tone in the muscles that close the eyes and in the elevator muscles of the lower jaw, and the sphincters of the rectum and bladder are normally contracted. The eyeballs are turned upward and outward, and the pupils are contracted (myosis) but respond to light. The loss of muscle tone and of postural reflexes causes the subject to fall if a recumbent position

has not been adopted before sleep. In some species, *e.g.*, the horse, postural reflexes are usually not suppressed during sleep; in man they may persist in exceptional conditions, *e.g.*, in persons who fall asleep while driving a car and in sleepwalking. Spontaneous movements are considerably reduced but not completely abolished during sleep; movements decrease as the depth of sleep increases. During sleep, abnormal motility, *e.g.*, tics, chorea, athetosis, myoclonia, etc., and certain other muscular rigidities, such as those of Parkinson's disease, hemiplegia, etc., decrease considerably or are abolished.

Reflex activity. Tendon reflexes diminish together with muscle tone, and some of them, *e.g.*, the knee jerk, are completely suppressed. The threshold of cutaneous reflexes is raised proportionally to the depth of sleep. Stimulation of the plantar surface provokes flexion of the toes in light sleep and extension (positive Babinski) in deep sleep; sometimes there is a mass flexion reflex.

Sensory activity. There is marked depression of sensory perception, especially of smell and taste, and to a lesser degree ot touch and hearing. There are considerable differences in the depression of the senses during sleep; thus a mother will quickly waken if her child cries or moves, but may not respond to much stronger stimuli of another nature.

Circulation. The heart rate diminishes by 10 to 30 beats per minute and the systolic pressure by 10 to 20 mm. Hg. The decrease is greater as the depth of sleep increases. Dreams and nightmares may provoke tachycardia and a considerable increase in blood pressure.

Respiration. The respiratory rate is usually decreased; inspiration is prolonged (often there is snoring) and expiration is brief. In light sleep the respiratory movements are rhythmic, but in deep sleep they often become irregular and periodic. Pulmonary ventilation decreases 25 per cent, therefore CO_2 partial pressure in alveolar air and blood is usually slightly higher than when the subject is awake.

Digestion and secretion. Gastric motility continues and may increase in vigor, and there is little or no change in gastric secretion. Stomach-emptying time and digestion are not influenced by sleep. Salivary and lacrimal secretions diminish considerably. The urine becomes more highly concentrated and is eliminated in smaller volume per unit period. Urinary excretion of chloride and nitrogen diminishes and that of phosphate increases. Sweat secretion increases considerably.

Metabolism. The BMR diminishes 10 to 15 per cent during deep sleep. The body temperature drops a few tenths of a degree, and is usually at a minimum between 2:00 A.M. and 4:00 A.M.

Electrolytes. There are no significant changes in the concentration of electrolytes in the blood. Hypnotics and anesthetics cause a fall in the concentration of K in plasma (Marenzi and Gerschmann)—not an increase, as was reported by Cloetta. There is no evidence that sleep is due to an increase in Ca in the diencephalon and that awakening is caused by an increase in K (Demole).

Humoral factors. According to Pieron activity provokes the accumulation of metabolites which produce somnolence and sleep; if animals are prevented from sleeping, a "hypnotoxin" appears in the blood which produces sleep. Humoral factors cannot, however, be the sole, or even the main cause of sleep, because one of a pair of Siamese twins may be asleep while the other is awake.

Cortical activity. The activity of the cerebral cortex is diminished in sleep but not totally abolished. Mental functions are depressed, dissociated, and in part disintegrated. The EEG shows marked and typical changes during sleep (see Chap. 87).

Sleep requirement. During sleep the organism is in a condition of mental and bodily rest propitious for the restoration of energy, especially that of the nervous system and in particular of the cortex. If dogs are deprived of sleep, they die after 12 to 17 days; occasionally animals resist as long as 3 weeks. Puppies cannot survive sleeplessness for more than 4 to 6 days. Cytologic changes (chromatolysis, cell shrinking, etc.) are observed. A few hours of sleep are sufficient to repair the effects of prolonged insomnia.

In man temporary deprivation of sleep does not produce any serious disturbance. Kleitman, Tyler, and others have made careful observations in normal subjects who were kept awake for periods of 60 to 200 hr. No significant cardiac, respiratory, or biochemical changes were observed. Mental processes were apparently normal. The subjects were nevertheless irritable, and the threshold for pain was diminished, but the threshold for touch was increased. The right-

ing reflexes were not as efficient as normally (there was difficulty in standing with the eyes closed), attention was sustained only with difficulty, and there was marked indifference to environmental stimuli; occasionally hallucinations occurred. The subjects easily fell asleep if they were not being constantly stimulated.

THE MECHANISM OF SLEEP

Electroencephalography has shown that there are different levels of general cortical activity. In somnolence and in sleep there is progressive synchronization of the electrical activity of the neurons, and large slow waves appear in the EEG. The awakening reaction is accompanied by desynchronization of the electrical activity of the neurons and the establishment of a rapid rhythm of small waves (see Chap 87). Sleep is fundamentally a cortical phenomenon, but the role of subcortical centers in sleep was already postulated by Mauthner (1890) many years ago. Later, the study of lesions found in the brain stem in cases of encephalitis or tumors, in which lethargy, somnolence or prolonged sleep, or insomnia had been an outstanding feature, showed that diencephalic and mesencephalic centers played an important part in the regulation of sleep and waking.[1]

Sleep, or the opposite, an increase in cortical activity, has been provoked by experimental stimulation or destruction of limited areas of the diencephalon and the bulbomesencephalic reticular formation. Hess implanted protected electrodes into the diencephalon of cats, and when the animals had recovered from the operation and were awake, stimulation provoked adynamia, loss of postural tone, and a state similar to sleep. Stimulation of the posterior hypothalamus, on the contrary, increased general activity (dynamogenic area). These results have been confirmed[2] and it has been shown that stimulation with low voltage and low frequency (3 to 5 per second), with electrodes implanted in the internal medullary lamina of the thalamus, provokes in waking animals the somatic signs of sleep, *i.e.*, muscular relaxation, typical posture of sleep, slow deep breathing, myosis,

[1] Von Economo, C., in Bethe's "Handbuch der normalen und pathologischen Physiologie," **17**, 591, Springer, Berlin, 1926; *J. Nerv. & Ment. Dis.*, **71**, 249, 1930.

[2] Akert, K., W. P. Koella, and R. Hess, Jr., *Am. J. Physiol.*, **168**, 260, 1952.

and the EEG of sleep (large, slow waves). Stimulation at higher frequency (10 to 30 per second) causes abrupt blocking of all spontaneous electrical activity of the cortex. High-frequency stimulation (200 per second) produces electroencephalographic and somatic signs of cortical activation; in some cases the animals enter into the condition called "sham rage" (see "Emotional expression," under "The hypothalamus," in Chap. 84).[1]

Stimulation of the caudal hypothalamus, or of the reticular formation in all its length from the roof of the mesencephalon to the medulla, produces signs of cortical activation in animals anesthetized with dial or chloralose and in Bremer's "isolated-encephalon" preparation: the slow large waves and sleep spindles are replaced by rapid small waves.[2]

Bremer[3] has shown that, in cats, section of the brain stem at the level of the upper part of the mesencephalon (*cerveau isolé*) provokes somatic signs of sleep in the head (myosis, closing of the eyes) and an EEG typical of sleep. If the section is made more caudally so that the cortex is connected with the mesencephalon (*encephale isolé*), the aspect of the head and the EEG are those of a waking animal. In the rat a complete transverse section immediately caudal to the hypothalamus produces a condition of permanent somnolence;[4] the animals can be wakened only by means of strong stimuli, and they fall asleep soon after stimulation ceases. The EEG is typical of sleep, and stimulation of afferent paths provokes signs of activation (desynchronization of the electrical activity of cortical neurons) only while stimulation lasts.

The state of wakefulness has been attributed to the continuous arrival of sensory impulses at the cortex (Bremer), especially those from proprioceptive receptors (Kleitman). It is well known that a decrease in afferent impulses caused by fatigue, muscular relaxation, or the suppression of auditory, visual, and other sensory stimuli tends to cause somnolence, but this

[1] Hunter, J., and H. H. Jasper, *Electroencephalog. & Clin. Neurophysiol.*, **1**, 305, 1949.

[2] Morrison, R. S., and E. W. Dempsey, *Am. J. Physiol.*, **135**, 281, 293, and 301, 1942; Moruzzi, G., and H. W. Magoun, *Electroencephalog. & Clin. Neurophysiol.*, **1**, 455, 1949.

[3] Bremer, F., *Compt. rend. Soc. de biol.*, **118**, 1241, 1935.

[4] Ingram, W. R., R. W. Barris, and S. W. Ranson, *Arch. Neurol. & Psychiat.*, **35**, 1175, 1936; Ranson, S. W., *Arch. Neurol. & Psychiat.*, **41**, 1, 1939.

is by no means the only or the main cause of sleep, because a subject may remain awake even when sensory stimulation has been reduced to a minimum.

Pavlov[1] attributed sleep to internal inhibition spreading to the whole cortex. He observed that in dogs summation of stimuli inhibitory for conditioned reflexes depressed all cortical activity and the animals appeared to fall asleep. Rhythmic repetition of a stimulus of low intensity (a monotonous voice, the hum of a motor, etc.) produces somnolence. Conditioned reflexes, such as the habit of retiring at a certain time, also produce sleepiness. Cortical activity can, therefore, be depressed by certain stimuli, just as it is increased by others.

General cortical activity is regulated by impulses arising in subcortical centers, but there is no evidence of a localized center producing sleep or waking. There is ample evidence that ascending stimuli from the reticular formation[2] and the caudal hypothalamus are necessary for the maintenance of waking, because their suppression causes lethargy or permanent somnolence, and stimulation of these areas increases cortical activity. Perhaps impulses from the anterior diencephalon depress cortical activity (Nauta). Impulses from intralaminar thalamic nuclei, the centrum medianum, etc., which give rise to the diffuse thalamic projection system, can inhibit or increase cortical activity according to the conditions of stimulation. Probably this system "may be involved in the 'higher level' integration of cerebral activity related to mechanisms of attention, consciousness and general alertness" (Hunter and Jasper). To sum up, sleep consists in the reversible depression of the activity of the whole cortex—activity which is controlled by impulses arising or integrated in the mesencephalon and diencephalon.

[1] PAVLOV, I. P., "Conditioned Reflexes," Oxford, New York, 1927.

[2] MAGOUN, H. W., and R. RHINES, *J. Neurophysiol*, **9**, 155 and 212, 1946.

BIBLIOGRAPHY

BREMER, F., Le Problème physiologique du sommeil, *Medicina, Parma*, **5**, 589, 1951.

HESS, W. R., Le Sommeil (Réunion plenaire), *Compt. rend. Soc. de biol.*, **107**, 1333, 1931.

HESS, W. R., "Der Methodik der lokalizierten Reizung," Thieme, Leipzig, 1932.

HESS, W. R., *Arch. f. Psychiat.*, **104**, 548, 1936.

HESS, W. R., "Zwischenhirn Syndrom," Schwabe, Basel (Grune and Stratton, New York), 1949.

HESS, W. R., Le Sommeil comme fonction physiologique, *J. physiol., Paris*, **41**, 61A, 1949.

KAYSER, C., Le Sommeil, *J. physiol., Paris*, **41**, 1A, 1949.

KAYSER, C., Le Sommeil hibernal, *Biol. Rev.*, **25**, 255, 1950.

KLEITMAN, N., *Physiol. Rev.*, **9**, 624, 1929.

KLEITMAN, N., "Sleep and Wakefulness," University of Chicago Press, Chicago, 1939.

KLEITMAN, N., in Glasser, "Medical Physics," vol. 1, p 1433. Year Book Publishers, Chicago, 1944.

Index

D

F

G

H

I

P

Q

R

S

T

U